Exploring social aspects in multilingual NLP:

(1) code-switching dialogues and (2) accent classification

with Emily Ahn

16 Oct 2020

NOTE: This talk only covered some of Part 1, and only those slides are included here.

Goals

- What aspects of Multilingual Language Technologies
  - are sociolinguists interested in?
  - should computational people be concerned about or aware of?
- How can we have dialogue across these domains?

Projects

- Code-switching in Human-Computer Dialogues
  - Spanish-English (Ahn et al., 2020)
  - Hindi-English (Parekh et al., 2020)
- Accent/dialect classification with Machine Learning

Code-switching in Human-Computer Dialogues

What Code-Switching Strategies are Effective in Dialogue Systems?

1 Carnegie Mellon University

2 University of Pittsburgh

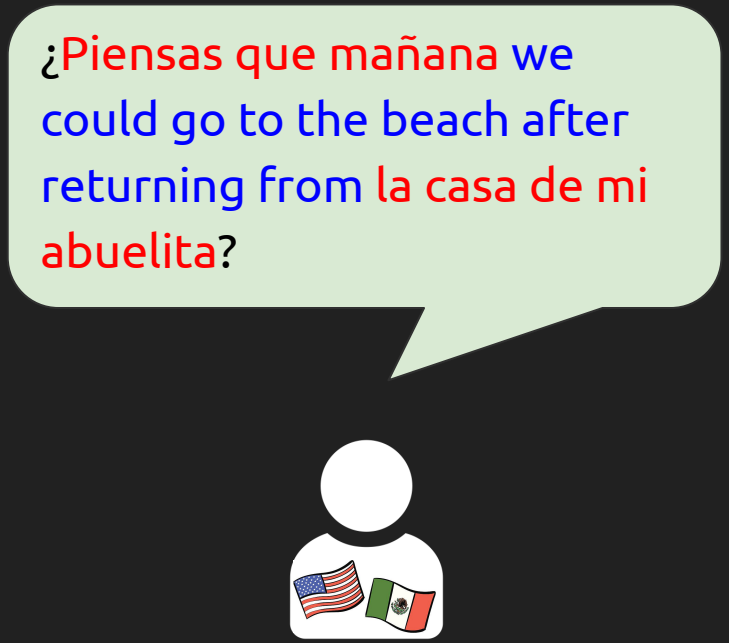
3 UNIVERSITY of WASHINGTON

SCiL
4 January 2020

Emily Ahn[3], Cecilia Jimenez[2], Yulia Tsvetkov[1], Alan Black[1]

Code-switching (CS)

Using multiple languages to communicate



[English]:

"Do you think that tomorrow we could go to the beach after returning from my grandmother's house?"

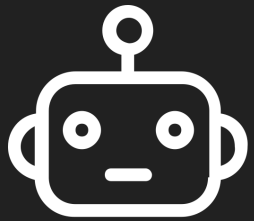
(Ardila, 2005)

Monolingual English Dialogue System

Do you have any friend who studies linguistics?

No, I only have friends who study computer science.

Why not both?

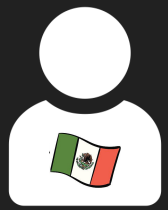


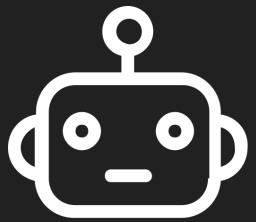
Monolingual Spanish Dialogue System

¿Tienes algún amigo que estudie lingüística?

No, solo tengo amigos que estudian Ciencias de la Computación.

¿Por qué no los dos?



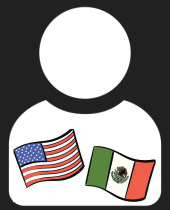


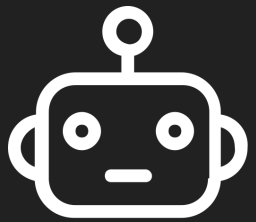
Bilingual Spanish-English Dialogue System: *Inserting English Content*

¿Tienes algún friend que estudie linguistics?

No, solo tengo friends que estudian Computer Science.

¿Por qué no los dos?



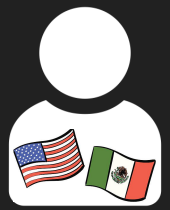


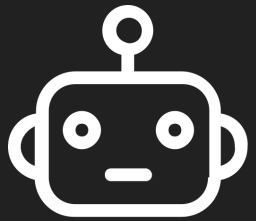
Bilingual Spanish-English Dialogue System: *Alternating Structure*

¿Tienes algún amigo that studies linguistics?

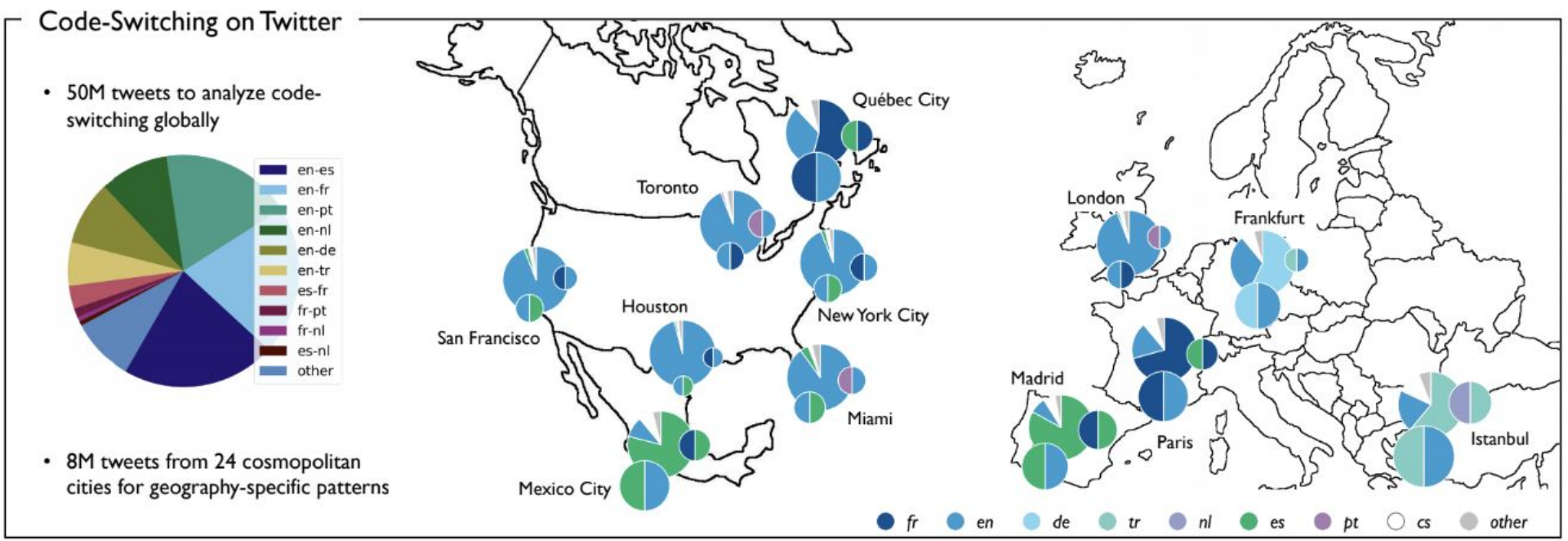
No, solo tengo amigos that study Computer Science.

¿Por qué no los dos?





Code-switching is a worldwide phenomena

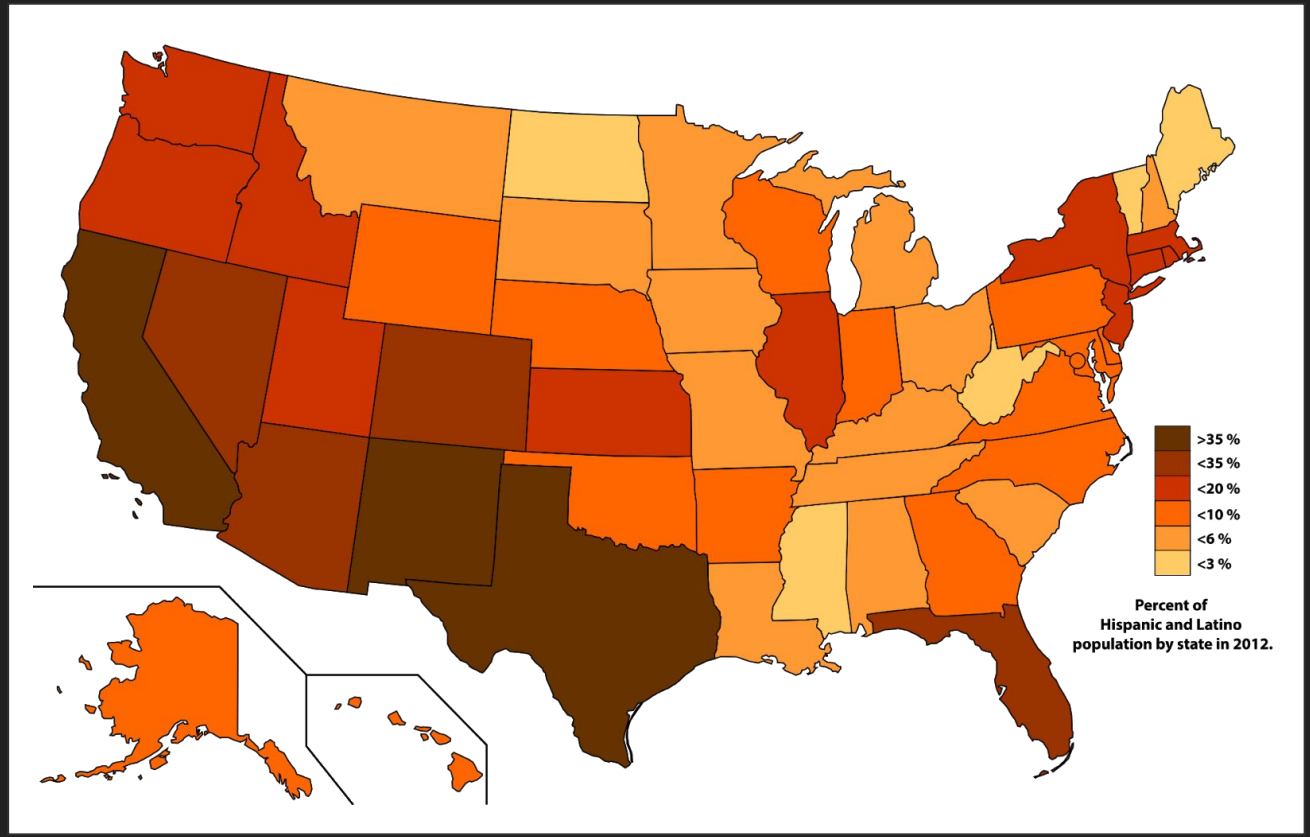


(Rijhwani et al., 2017)

Spanish-English code-switching is common in the US

Hispanic population
~ 18% total US population
(US Census Bureau, 2017)

(https://en.wikipedia.org/wiki/Demographics_of_Hispanic_and_Latino_Americans)



Motivations

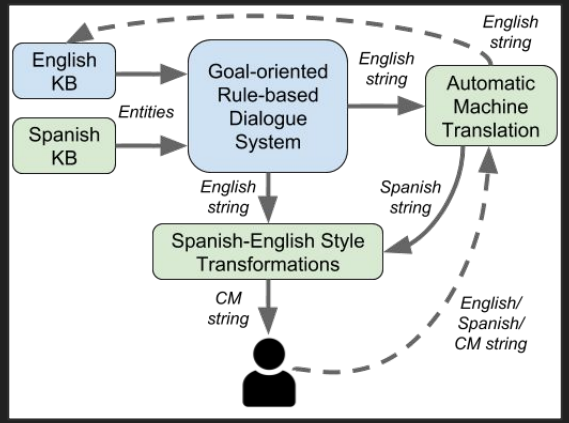
Improve future of multilingual, human-centered NLP technology

Towards future code-switching human-computer dialogue systems

Learn human code-switching preferences within human-computer dialogues

Our 3 Contributions

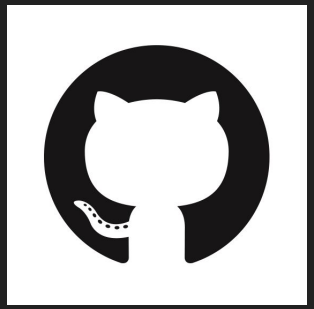
[C1] New framework of incorporating code-switching strategies into a bilingual collaborative dialogue system



Our 3 Contributions

[C2] New corpus of 587 code-switched Spanish-English human-computer text dialogues

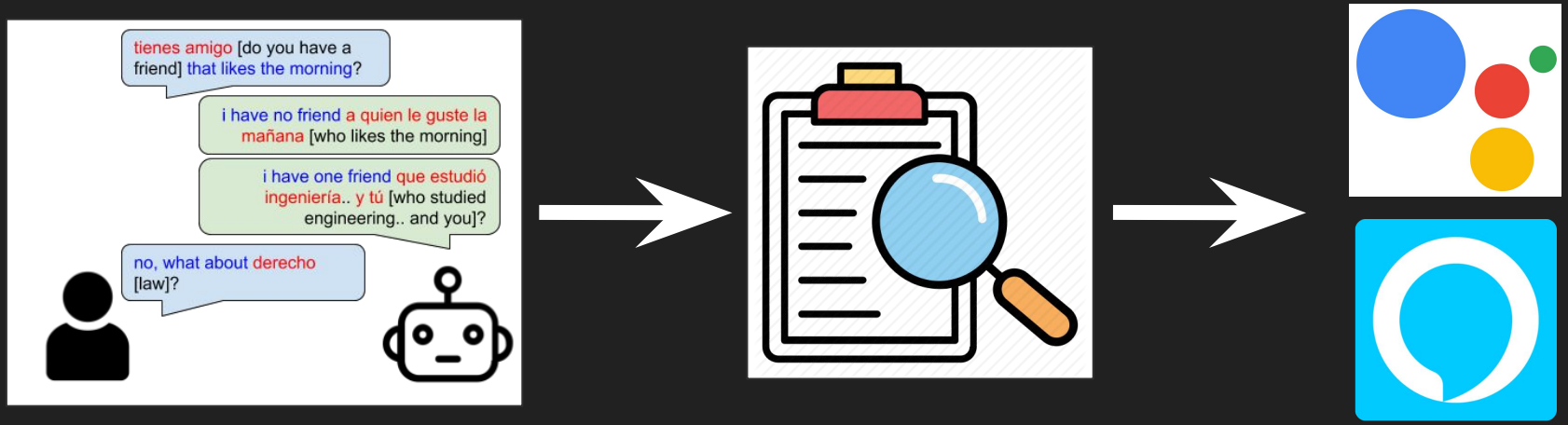
COMMONAMIGOS



https://github.com/emilyahn/commonamigos

Our 3 Contributions

[C3] Exploratory analyses of code-switched dialogues/strategies, and recommendations for future bilingual dialogue systems



Outline

1. Background
2. [C1] Our code-switching dialogue system
3. [C2] Data collection
4. [C3] Analysis
5. Conclusion

Background

1. Background
2. Our code-switching dialogue system
3. Data collection
4. Analysis
5. Conclusion

Prior Work

- Code-switching research: Trying to predict CS switch-points
  - Acoustically (Voigt et al., 2016)
  - Syntactically (Solorio & Liu, 2008)
  - Via cognates (Kootstra+ 2012)
- NLP/Speech applications
  - Language ID (Rijhwani+ 2017)
  - Automatic Speech Recognition (Yilmaz et al., 2018)
- Spanglish linguistics
  - Cultural overview (Ardila, 2005)
  - Formal grammar (Sankoff & Poplack, 1981)
- Entrainment
  - Rapport influences style of CS (Jarvis & Pavlenko, 2006)
  - Linguistic/phonetic (Fricke+ 2016)

Existing code-switching system: fixed utterances

English	English–Hindi	English–Spanish
Hi, welcome to The Coffee Spot. What can I get you today?	Hi! Coffee Spot me aapka swaagat hai! Would you like something to drink?	Hola! Bienvenido al Coffee Spot. Would you like something to drink?
Okay, Is that for here or to go?	Achha, got it. Would you like it for here ya phir parcel lenge?	Muy bien, I got it. Would you like it for here o para llevar?
Okay, would you like that hot or iced?	And would you like that thanda ya garam?	And would you like that frio o caliente?
And did you want that drink to be small, medium, or large?	OK. Aur aapko small, medium ya large size chahiye?	OK. Que tamaño lo quiere, small, medium or large?
And would you like that with milk or sugar/cream?	And would you like that with doodh ya cheeni?	And would you like that con crema o leche?
Perfect. Did you want something to eat with that?	Theek hai, perfect. Aur aapko kuch aur chahiye tha, to eat with that?	Muy bien, perfect. Le gustaria algo mas de comer, with that?
And I'm assuming you'd like that toasted?	And I'm assuming ki aapko woh toasted chahiye, right?	And I'm assuming le gustaria tostado?
Alright, thanks. Your order will be out shortly.	Okay ji. Your order will be ready shortly.	Gracias. Your order will be ready shortly.

Ramanarayanan & Suendermann-Oeft (2017)

Research Questions

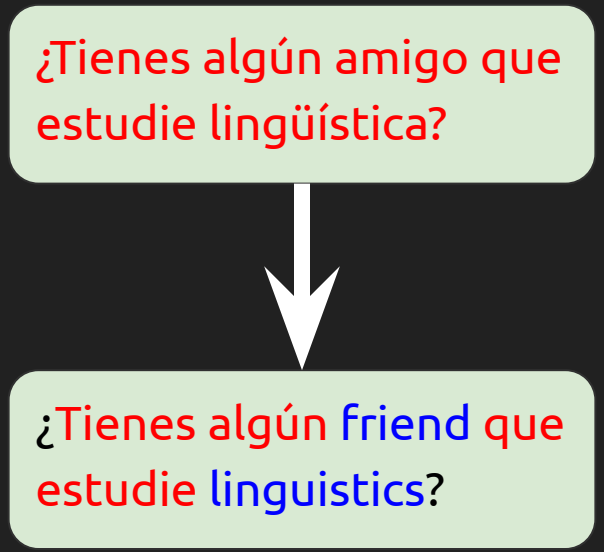
1. Can we elicit code-switching? How much?
2. What strategies do people use? Do users entrain to the agent?
3. Are there correlations with dialogue success?
4. Does gender have any effect?
5. Does language proficiency have any effect?

Code-switching Strategies

Strategy 1: Insertional CS

Replace content words with Embedded Language

(Muysken, 2000)



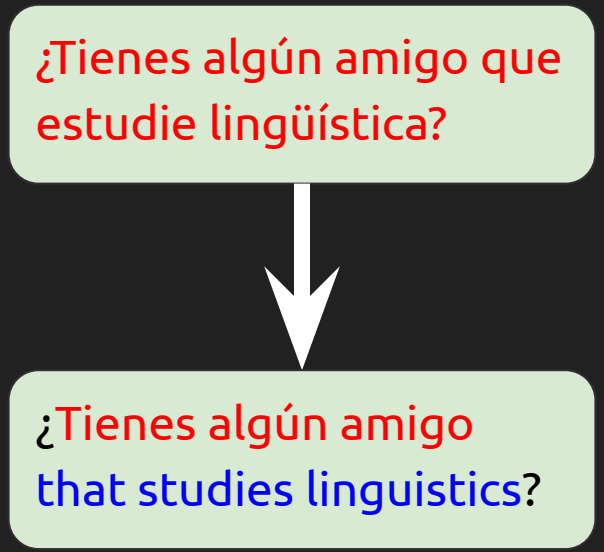
[English]:

“Do you have any friend that studies linguistics?”

Strategy 2: Alternational CS

Switch Matrix Language at syntactic breakpoint

(Muysken, 2000)



[English]:

“Do you have any friend that studies linguistics?”

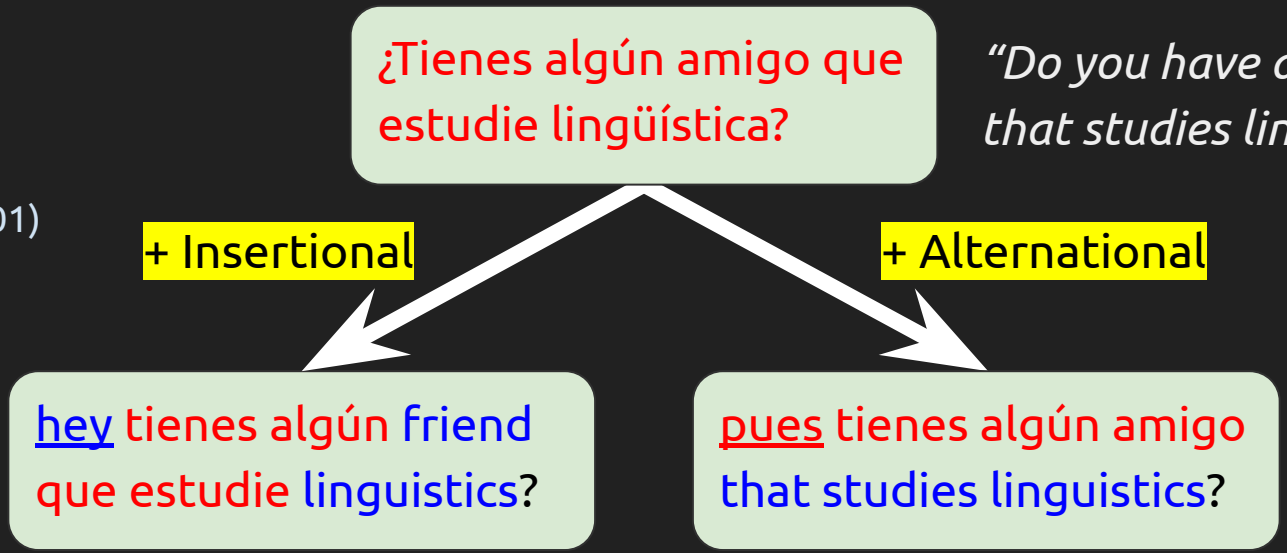
Strategy 3: Informal CS

Add discourse markers

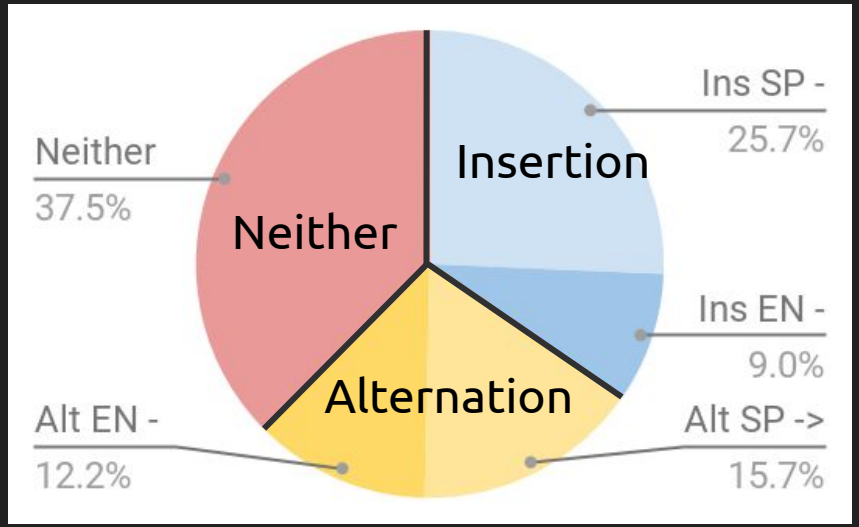
(Torres, 2001)

[English]:

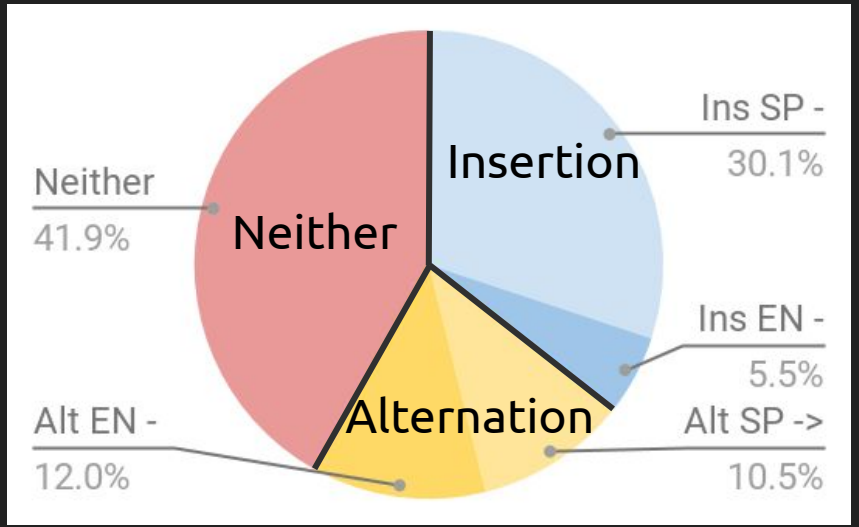
“Do you have any friend that studies linguistics?”



Validation: these strategies occur in other corpora



Miami Bangor corpus
(spontaneous speech)
(Deuchar et al., 2014)



Twitter corpus
(text)
(Molina et al., 2016)

[C1] Our code-switching dialogue system

1. Background
2. Our code-switching dialogue system
3. Data collection
4. Analysis
5. Conclusion

Existing monolingual goal-oriented dialogue system

→ Find your mutual friend

(He et al., 2017)

Your friends

# ▲▼	School ▲▼	Time Preference ▲▼	Location Preference ▲▼
Select	University of the Ozarks	afternoon	outdoor
Select	Christian Brothers University	afternoon	outdoor
Select	Radford University	afternoon	indoor
Select	North Carolina Wesleyan College	morning	outdoor

Time Remaining: 2:10

[02/09/18 00:17:18] <You entered the room.>
[02/09/18 00:17:19] Partner: hi
[02/09/18 00:17:28] Partner: I have 1 university of illinois at springfield, 1 radford university.
[02/09/18 00:17:44] You: sup
[02/09/18 00:17:50] Partner: Do you have any friends who like outdoor?
[02/09/18 00:17:57] You: i have a couple at radford

Enter your message here

[C2] Data collection

1. Background
2. Our code-switching dialogue system
3. Data collection
4. Analysis
5. Conclusion

Crowdsourcing dialogues

Screening crowdworkers:

- Location = USA only
- 3-question multiple-choice Spanish proficiency quiz
- Task title: "*Charlemos en Spanglish!*" [Let's chat in Spanglish!]
- Instructions mainly in Spanish

> *"You may write in English, Spanish, or a combination of the two"*

Bilingual interface

¿Quién es nuestro amigo en común?

Usted y otro usuario tienen exactamente un amigo en común. Usted sabe unos atributos para cada uno de sus amigos (como nombre, escuela, etc.). Tu meta es encontrar el amigo en común usando estos atributos!

Instrucciones

- Por favor use **oraciones naturales** lo más posible.
  - **Escribe:** tres de mis amigos trabajan en el banco
  - **No Escribas:** 3 banco
- Evite simplemente enumerar los atributos (nombre, compañía, etc.) de sus amigos.
- Mira **la lista de amigos** a la derecha.
- Usa **la ventana de chat** para descubrir más de los amigos de su pareja.

Tus amigos / Your friends

# ▲▼	Especialidad Major ▲▼	Hobby Hobby ▲▼	Tiempo Time ▲▼
Select	educación education	actuar acting	la noche night
Select	administración management	el ganchillo crocheting	la mañana morning
Select	diseño gráfico graphic design	correr running	la noche night

Time / Tiempo: 7:10

[06/22/18 17:30:02] <You entered the room.>
[06/22/18 17:30:03] <Your partner has joined the room.>

Enter your message here

Overview of our collected corpus

Demographics:

- 296 unique users
- 60% male
- Mean age = 31
- Most frequently reported countries of origin: USA, Venezuela, Mexico

# Dialogues	587
% Extrinsic task success	64%
Avg # user utterances	7.9
Avg # tokens / utterance	6.2
EN vocab size	571
SP vocab size	846
% EN utterances	16%
% SP utterances	44%
% CS utterances	39%
% dialogues w/ CS	70%

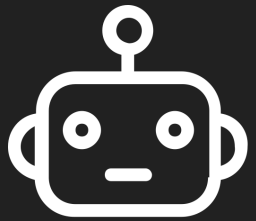
Sample from our corpus: Agent = Alternational CS

Do you have any friend who likes dancing
o amigos a los que les guste hornear?

[Do you have any friend who likes dancing or friends that like to bake]

[no one likes to dance. One likes baking--he/she studies physics]

nadie le guste bailar.
one likes baking--el/ella estudia fisica.



[C3] Analysis

1. Background
2. Our code-switching dialogue system
3. Data collection
4. Analysis
5. Conclusion

Research Questions

1. Can we elicit code-switching? How much?
2. What strategies do people use? Do users entrain to the agent?
3. Are there correlations with dialogue success?
4. Does gender have any effect?
5. Does language proficiency have any effect?

1. Yes, we can elicit CS!

M-idx = Multilingual Index
→ how balanced are # tokens in each language

I-idx = Integration Index
→ how often are switch-points within an utterance

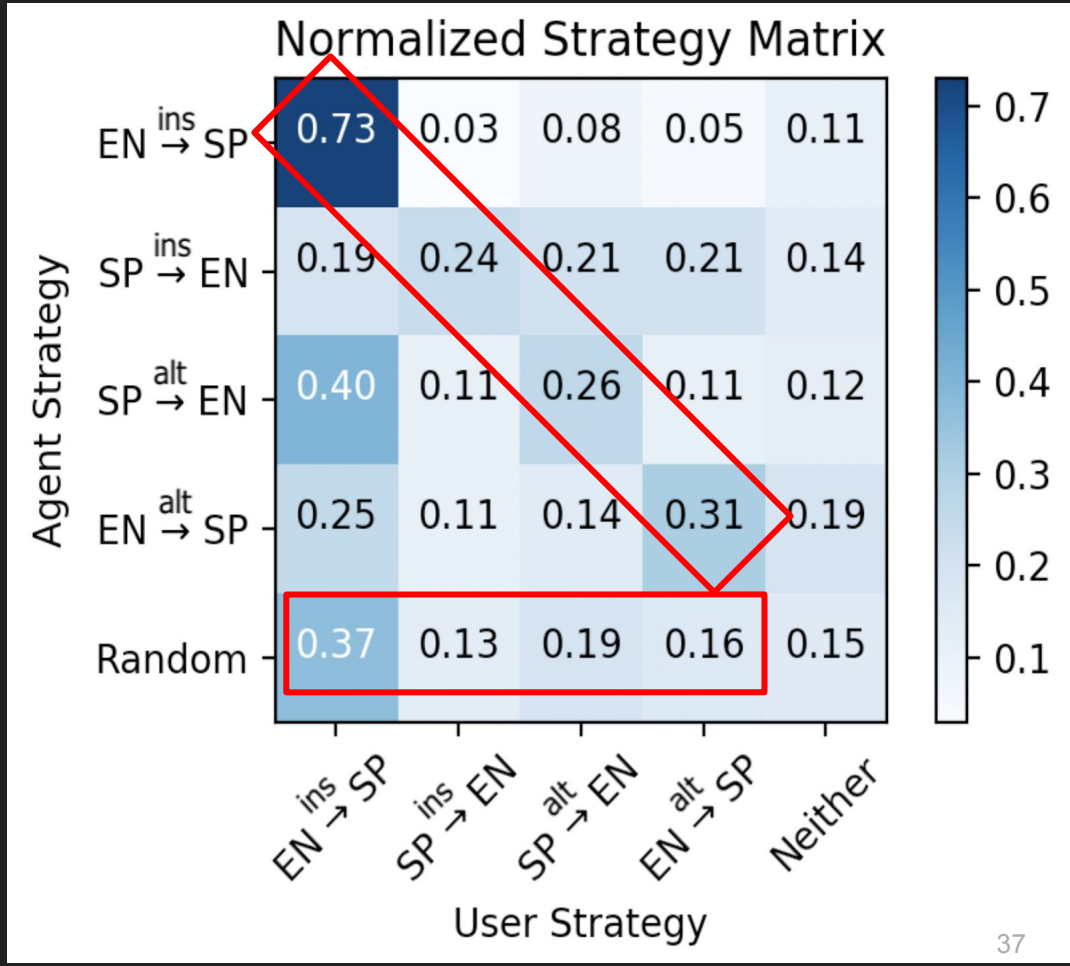
(Guzman et al., 2017)

Agent	% Dial. w/ CS	% Utts = CS	M-idx	I-idx
Average	70	39	0.74	0.23
Std Dev	(8)	(8)	(0.20)	(0.04)
$EN \xrightarrow{ins} SP$	74	42	**0.51**	0.23
+Informal	**80**	44	0.57	0.26
$SP \xrightarrow{ins} EN$	74	**52**	0.93	0.26
+ Informal	75	37	**0.99**	0.26

2. There is some strategy entrainment

Given all user's CS utts, what % are in each category?

- Perfect entrainment = dark diagonal lines
- User strategy in given condition is higher than baseline agent (Random)

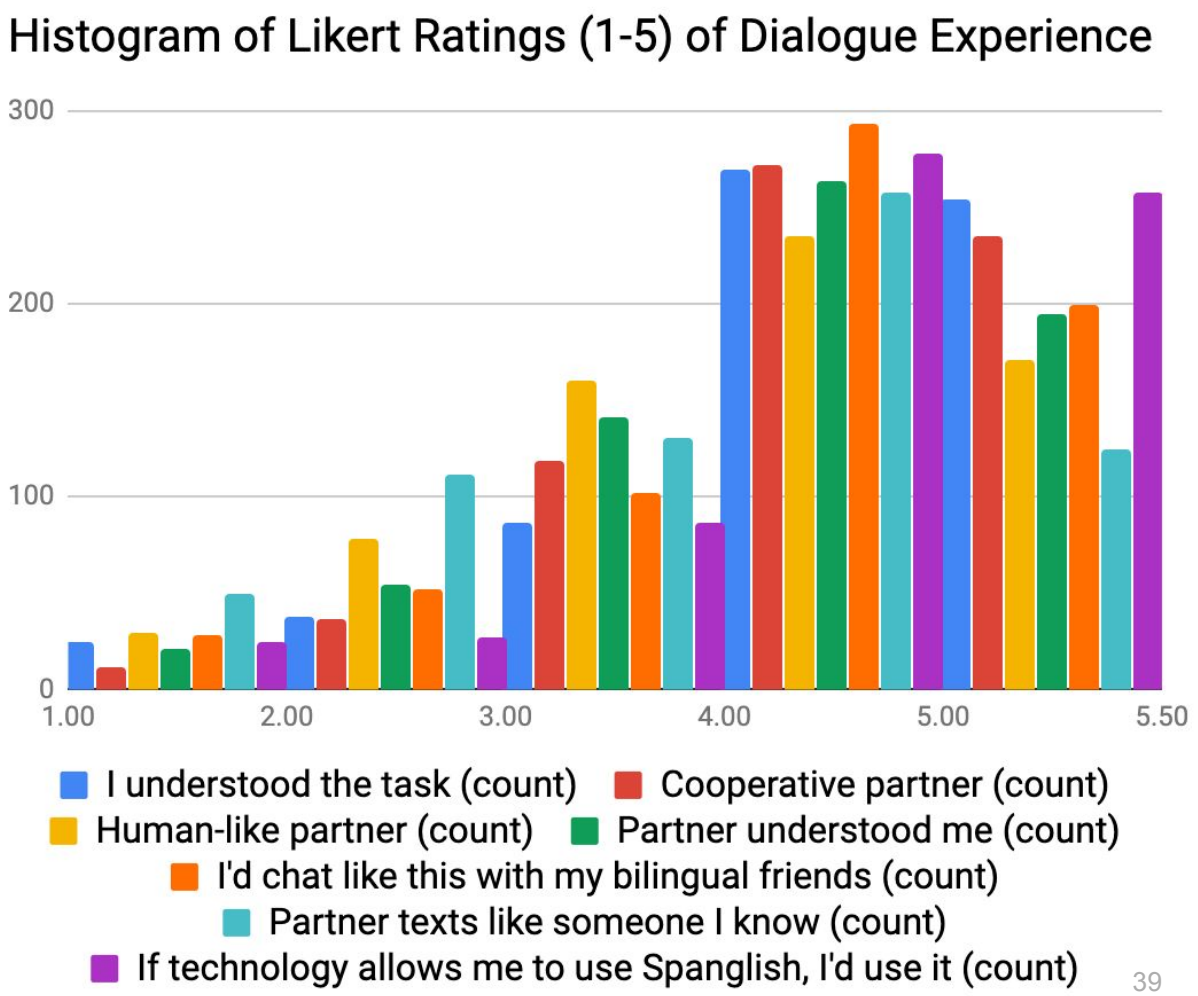


Part 2: Linguistic Background

1. At what age did you begin to learn Spanish?
2. At what age did you begin to learn English?
3. What is your current age?
4. What is your gender?
5. Rate your overall Spanish ability on a scale from 1-5.
6. Rate your overall English ability on a scale from 1-5.
7. Which country/countries do you or your family originate from?
8. What percentage of the time do you write online (i.e. text, social media, etc.) in Spanish?
9. What percentage of the time do you write online (i.e. text, social media, etc.) in English?
10. What percentage of the time do you write online (i.e. text, social media, etc.) in mixed Spanish and English?

3. Overall positive qualitative user experience

- User survey: Likert scale
  - 1 = Strong **Disagree**
  - 5 = Strong **Agree**



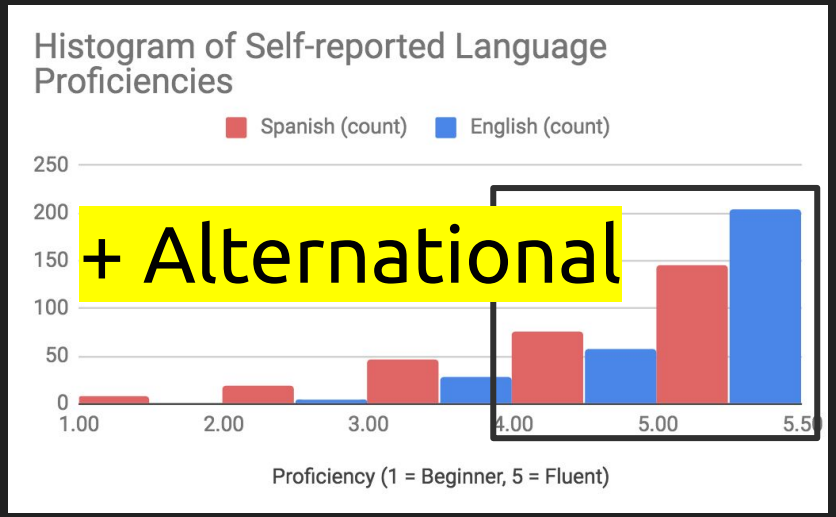
4. Females code-switched more than males in Informal conditions

- When agent used Informal strategy, females had
  - Longer dialogues
  - Higher quantity of CS

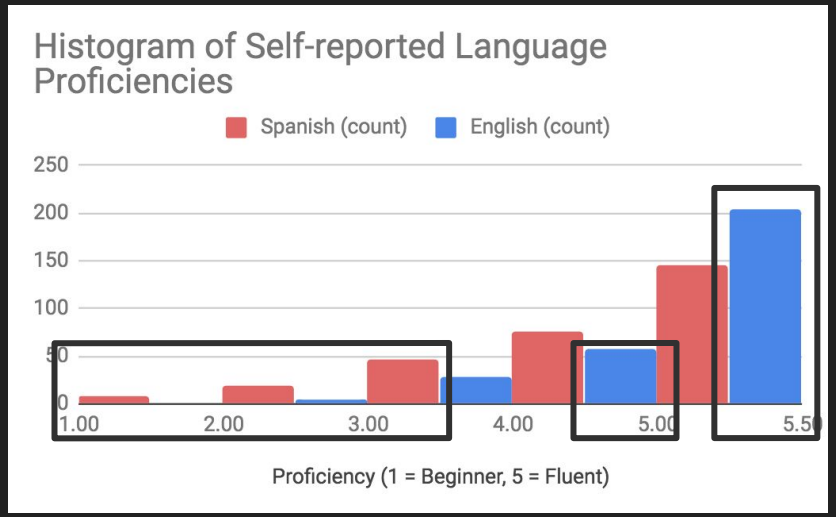
- When females CS more, they agreed more with:
  “I am very likely to chat like I did in this task when messaging with my bilingual friends” ($p < .005$)

5a. Verified linguistic hypothesis:

Symmetric bilinguals used Alternational CS more than asymmetric bilinguals (Deuchar et al., 2007)



Symmetric bilinguals



Asymmetric bilinguals

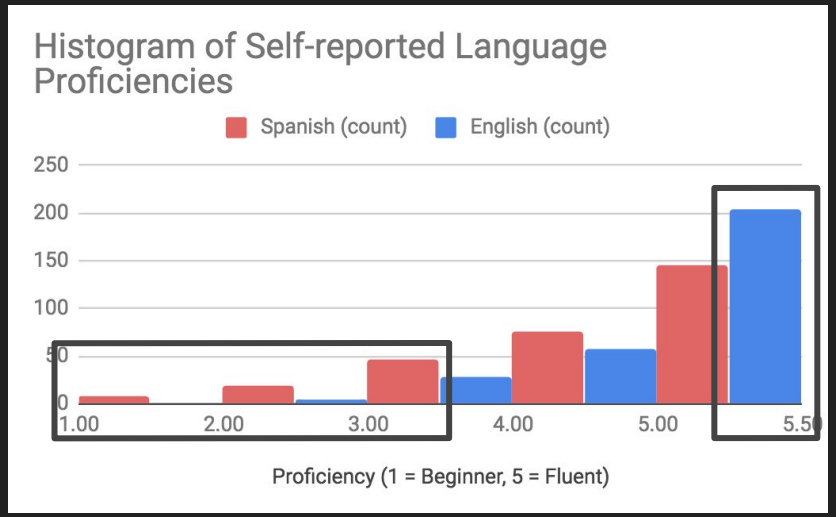
5b. Asymmetric bilinguals had better experience with CS agent than monolingual agent

For dominant English speakers:

CS agent vs. Spanish agent

+ Task Success
+ “cooperative partner”

Supports incorporating CS in Second Language Instruction
(cf. Moore, 2002)



Asymmetric bilinguals

Conclusion

1. Background
2. Our code-switching dialogue system
3. Data collection
4. Analysis
5. Conclusion

Summary of findings

1. We can elicit code-switching from users
2. There is some strategy entrainment
3. Users succeeded in their dialogues
4. Gender: females code-switched more than males in Informal conditions
5. Language proficiency:
    a. Symmetric bilinguals used Alternational CS more than asymmetric bilinguals
    b. Asymmetric bilinguals had better experience with CS agent than monolingual agent

Takeaways

- Important to consider strategies within CS
- There is much hope for code-switched conversational agents

Limitations & Future Work

- Population of crowdsourced workers
- Users’ attitudes towards CS?
- Should try other language pairs
- Agent should dynamically entrain to user
- Enhance dialogue system (not rule-based)

User Feedback

my partner was all over the place. It was hard to keep up with all their preguntas [questions]. I don't think we ever found our amigo en comun [mutual friend].

User Feedback

in my experience speaking with other bilingual friends the switching is usually either half of the sentence or alternating sentences...

Rarely do I just use one word in the other language unless it's pretty specific to that language...

But I found myself doing it along with the robot!

User Feedback

Felt that I was chatting con una persona real [with a real person]!

The team

Emily Ahn
University of Washington

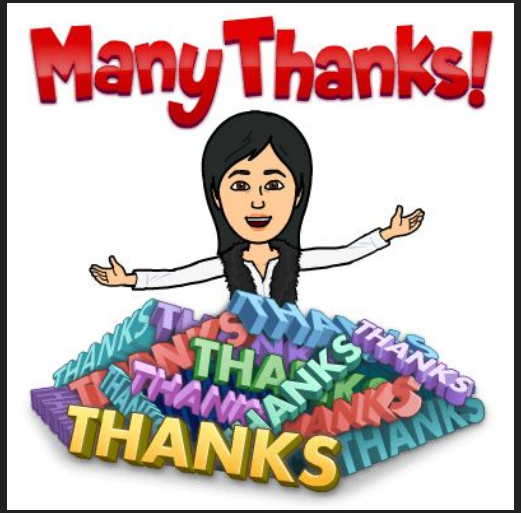
Cecilia Jimenez
University of Pittsburgh

Yulia Tsvetkov
Carnegie Mellon University

Alan W. Black
Carnegie Mellon University

References

1. Ardila, A. (2005). Spanglish: an anglicized Spanish dialect. *Hispanic Journal of Behavioral Sciences*, *27*(1), 60-81.
2. Deuchar, M., Muysken, P., & Wang, S. L. (2007). Structured variation in codeswitching: towards an empirically based typology of bi
3. Deuchar, M., Davies, P., Herring, J., Couto, M. C. P., & Carter, D. (2014). Building bilingual corpora. *Advances in the Study of Bilingualism*, 93-111.
4. Fricke, M., & Kootstra, G. J. (2016). Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, *91*, 181-201.
5. He, H., Balakrishnan, A., Eric, M., & Liang, P. (2017). Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *arXiv preprint arXiv:1704.07130*.
6. Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.
7. Kootstra, G. J., Van Hell, J. G., & Dijkstra, T. (2012). Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *Bilingualism: Language and Cognition*, *15*(4), 797-819.
8. Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., & Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching* (pp. 40-49).
9. Moore, D. (2002). Code-switching and learning in the classroom. *International journal of bilingual education and bilingualism*, *5*(5), 279-293.
10. Muysken, P., Díaz, C. P., & Muysken, P. C. (2000). *Bilingual speech: A typology of code-mixing* (Vol. 11). Cambridge University Press.
11. Ramanarayanan, V., & Suendermann-Oeft, D. (2017). Jee haan, I'd like both, por favor: Elicitation of a Code-Switched Corpus of Hindi–English and Spanish–English Human–Machine Dialog. *Proc. Interspeech 2017*, 47-51. Chicago
12. Rijhwani, S., Sequiera, R., Choudhury, M., Bali, K., & Maddila, C. S. (2017). Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1971-1982).\
13. Sankoff, D., & Poplack, S. (1981). A formal grammar for code-switching. *Research on Language & Social Interaction*, *14*(1), 3-45.
14. Solorio, T., & Liu, Y. (2008, October). Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 973-981). Association for Computational Linguistics.
15. Torres, L. (2011). Spanish in the United States: Bilingual discourse markers. *The Handbook of Hispanic Sociolinguistics*, 491-503.
16. Voigt, R., Jurafsky, D., & Sumner, M. (2016). Between-and Within-Speaker Effects of Bilingualism on F0 Variation. In *Interspeech* (pp. 1122-1126).
17. Yılmaz, E., Biswas, A., van der Westhuizen, E., de Wet, F., & Niesler, T. (2018). Building a Unified Code-Switching ASR System for South African Languages. *arXiv preprint arXiv:1807.10949*. Chicago



Questions?

eahn@uw.edu

https://www.seattletimes.com/sports/uw-huskies/10-10-would-cheer-with-uw-introduces-new-live-mascot-dubs-ii-and-he-is-adorable/