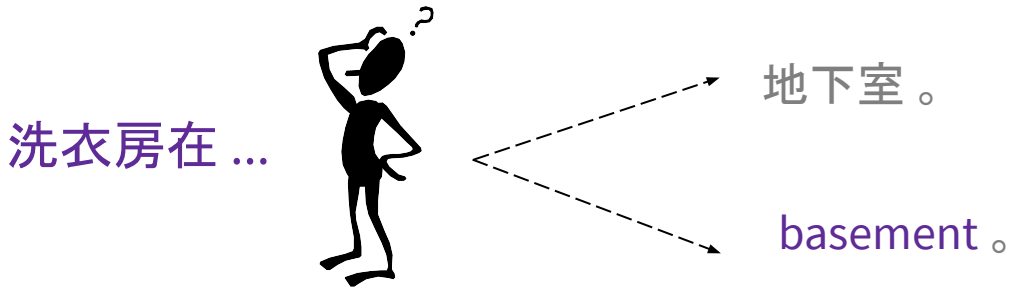
# The cross-lingual interaction that drives Chinese-English code-switching

Debasmita Bhattacharya

1. What is code-switching?
2. Related work
3. Hypothesis
4. Method
5. Preliminary Results
6. Upcoming plans

# What is code-switching?

- When a speaker alternates between one or more languages during linguistic communication
  - Can occur in written or spoken communication
  - Can involve switching between dialects of the same language
  - Very common among bilinguals!




(= The laundry room is in the basement.)

# Related work

- *Surprisal Predicts Code-Switching in Chinese-English Bilingual Text*
  - Calvillo et al.
  - EMNLP 2020
- *Simulating Spanish-English Code-Switching:* El Modelo Está *Generating Code-Switches*
  - Tsoukala et al.
  - ACL 2019
- *Simulating Code-switching Using a Neural Network Model of Bilingual Sentence Production*
  - Tsoukala et al.
  - Computational Brain and Behaviour 2021

# Related work: Calvillo et al.

**Surprisal Predicts Code-Switching in Chinese-English Bilingual Text**

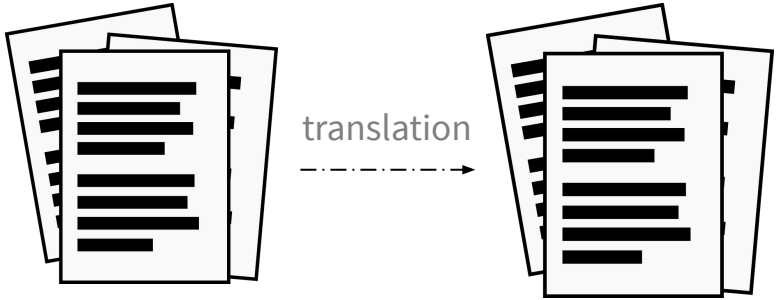
**Jesús Calvillo** [1] jzc1104@psu.edu
**Le Fang** [1] fredfang1203@gmail.com
**Jeremy Cole** [2,1] jrcole@google.com
**David Reitter** [2,1] reitter@google.com

[1]Pennsylvania State University
[2]Google Research

- Calvillo et al.: predictability of Chinese-English code-switching
  - Compared code-switched sentences to fully Chinese translations
    - “洗衣房在 basement” vs “洗衣房在 地下室”
- Chinese equivalent of code-switched word is “difficult” for a speaker to access
  - Modelled using surprisal

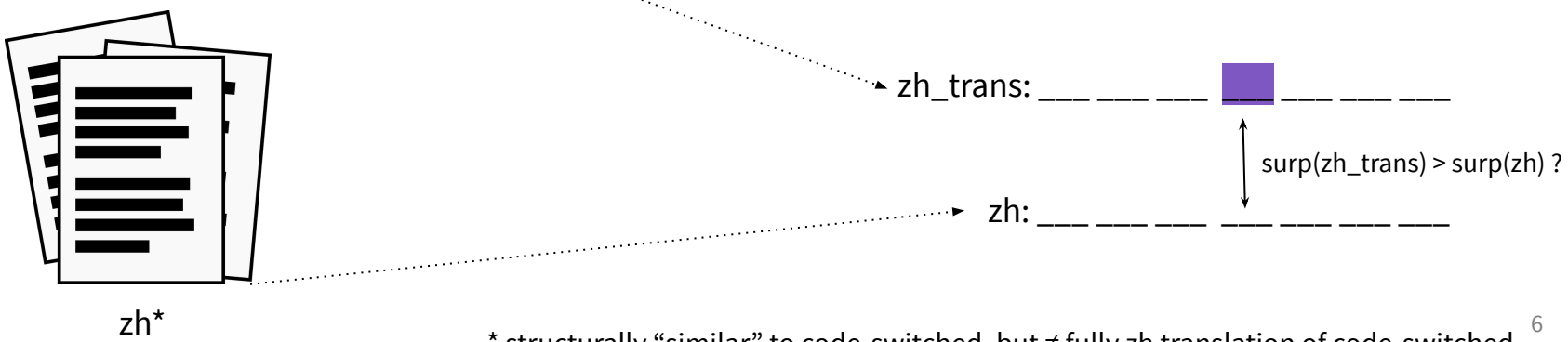
$$surp(w_i) = -logP(w_i|w_{i-1}, ..., w_{i-t})$$

# Hypothesis: Calvillo et al.



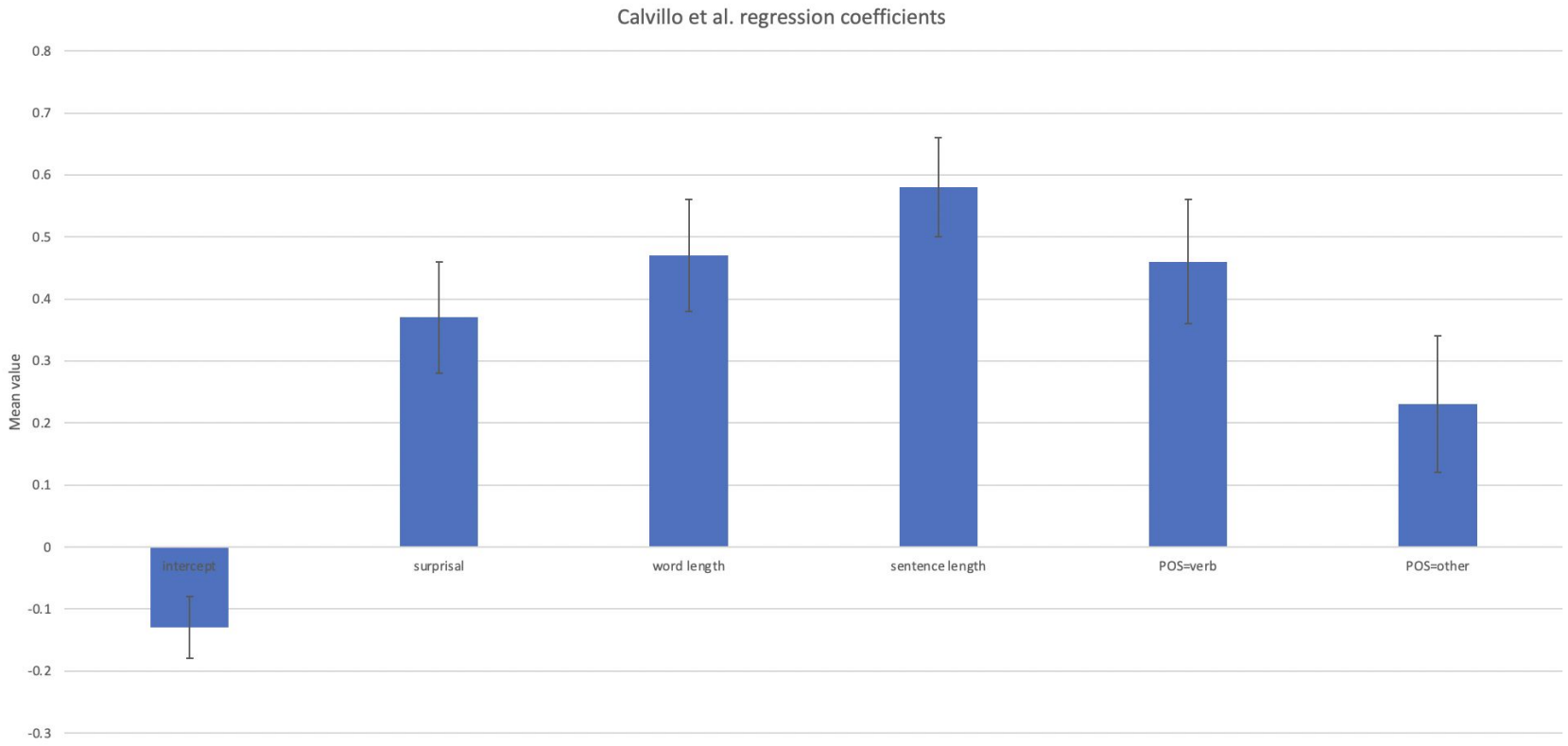

| | | | | | |
|---|---|---|---|---|---|
| original | 整个 | **house** | 家具 | | 齐全 |
| CS-sent | 整个 | **房子** | 家具 | | 齐全 |
| POS | DT | **NN** | NN | | VA |
| | whole | **house** | furniture | | complete |
| nonCS-sent | 全部 | 木头 | 地板 | , | 干净 |
| POS | DT | **NN** | NN | PU | VA |
| | all | **wood** | floor | | clean |

Table 1: Example of CS-sent / nonCS-sent alignment.

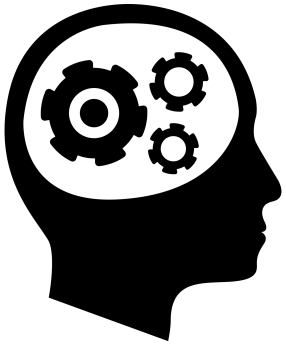



* structurally “similar” to code-switched, but ≠ fully zh translation of code-switched

# Results: Calvillo et al.



# My hypothesis



| | w1 | w2 | w3 |
|---|---|---|---|
| zh: | 洗衣房 | 在 | 地下室 |
| en: | The laundry room | is in the | basement |
| | surp(zh) < surp(en) | surp(zh) < surp(en) | surp(zh) > surp(en) |

# Method

- Data collection:
  - scraped 50K English Reddit posts + 50K Chinese Baidu Tieba posts
- Model training and testing:
  - Trained 5-gram models on (preceding *n-gram* + CS word)
  - Trained skip-gram models on (dep. relation + CS word + CS word dep. governor)
  - Tested on original Calvillo sentences

然后 就 可以 email 群发 附近 的 dealer 要求 报价，然后 让 他们 互相 beat 。

然后 就 可以 电子 邮件 群发 附近 的 卖家 要求 报价，然后 让 他们 互相 竞价 。

obj

code-switched word

preceding n-gram, n<=5

dependency governor

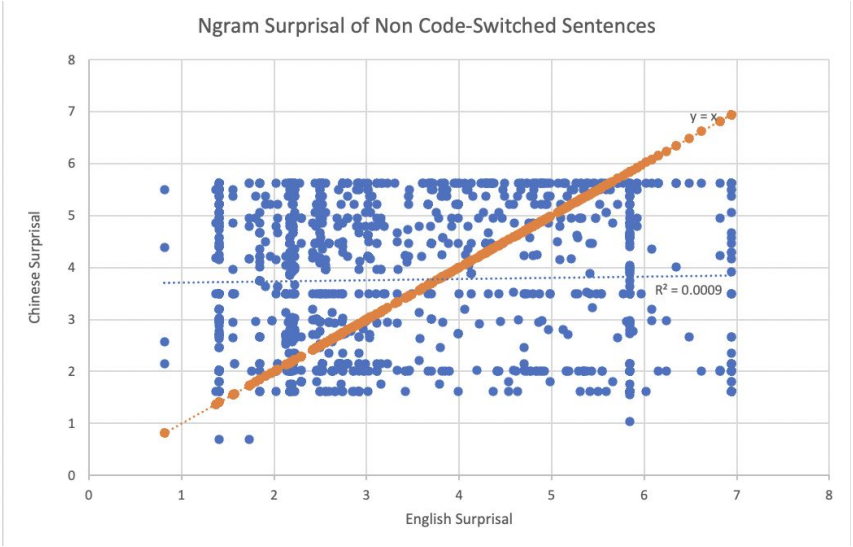
Then you can send emails to nearby dealers to ask for quotation, and then let them beat each other.

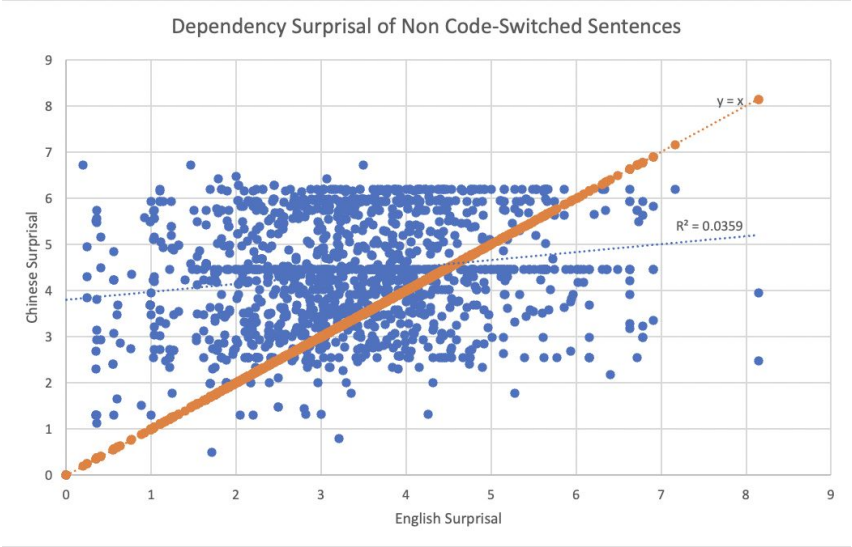
obj

# But… automatic translation??

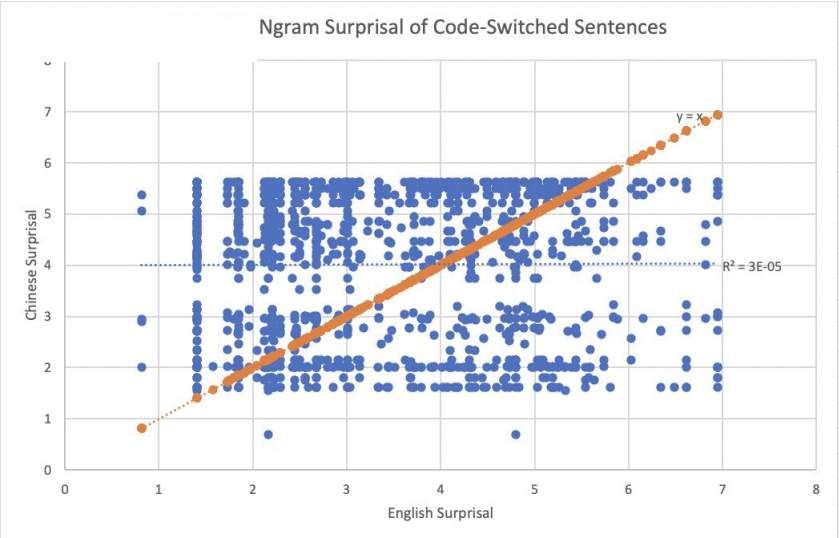
🤔🧐

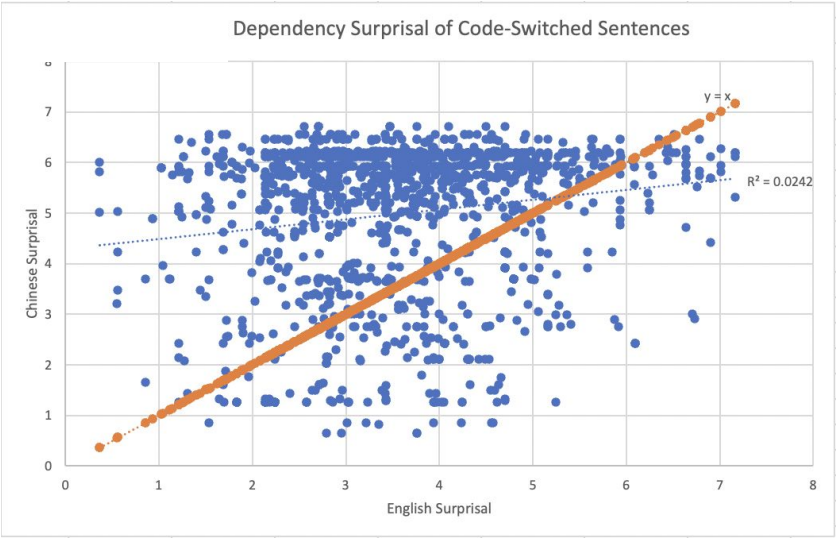
# Preliminary Results: control data



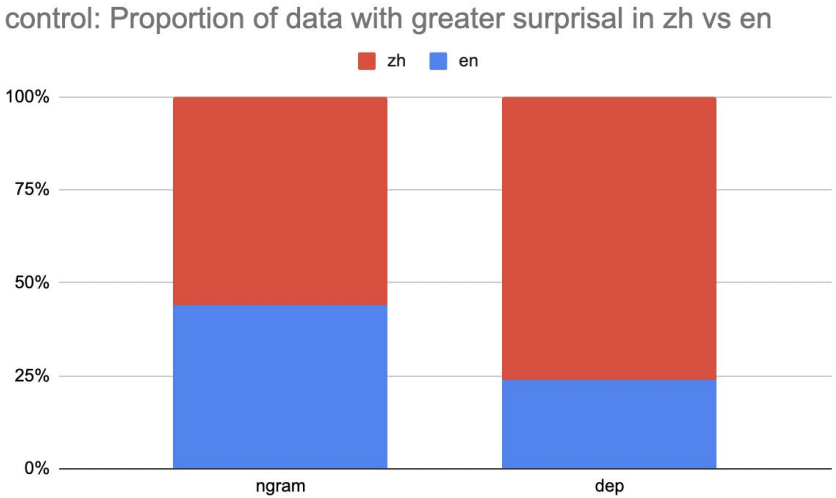



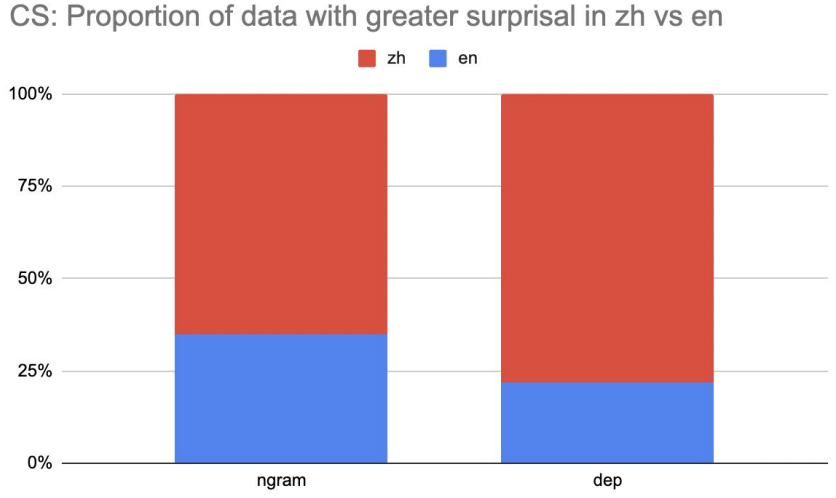
# Preliminary Results: code-switched data



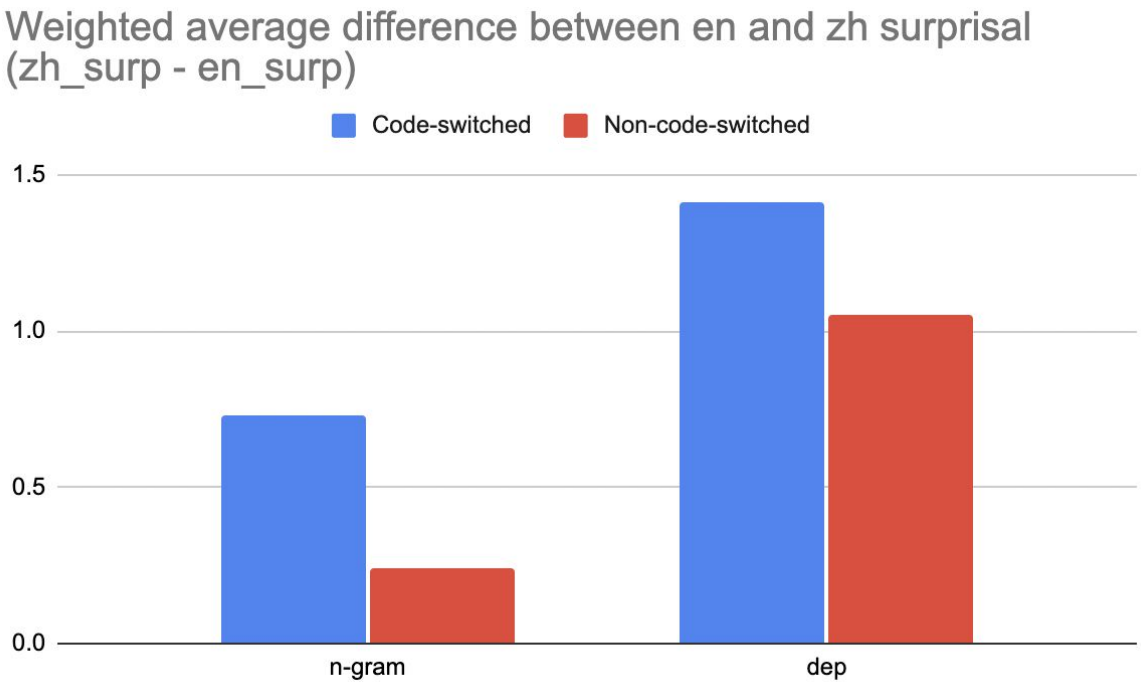



# Preliminary Results: comparing control to cs

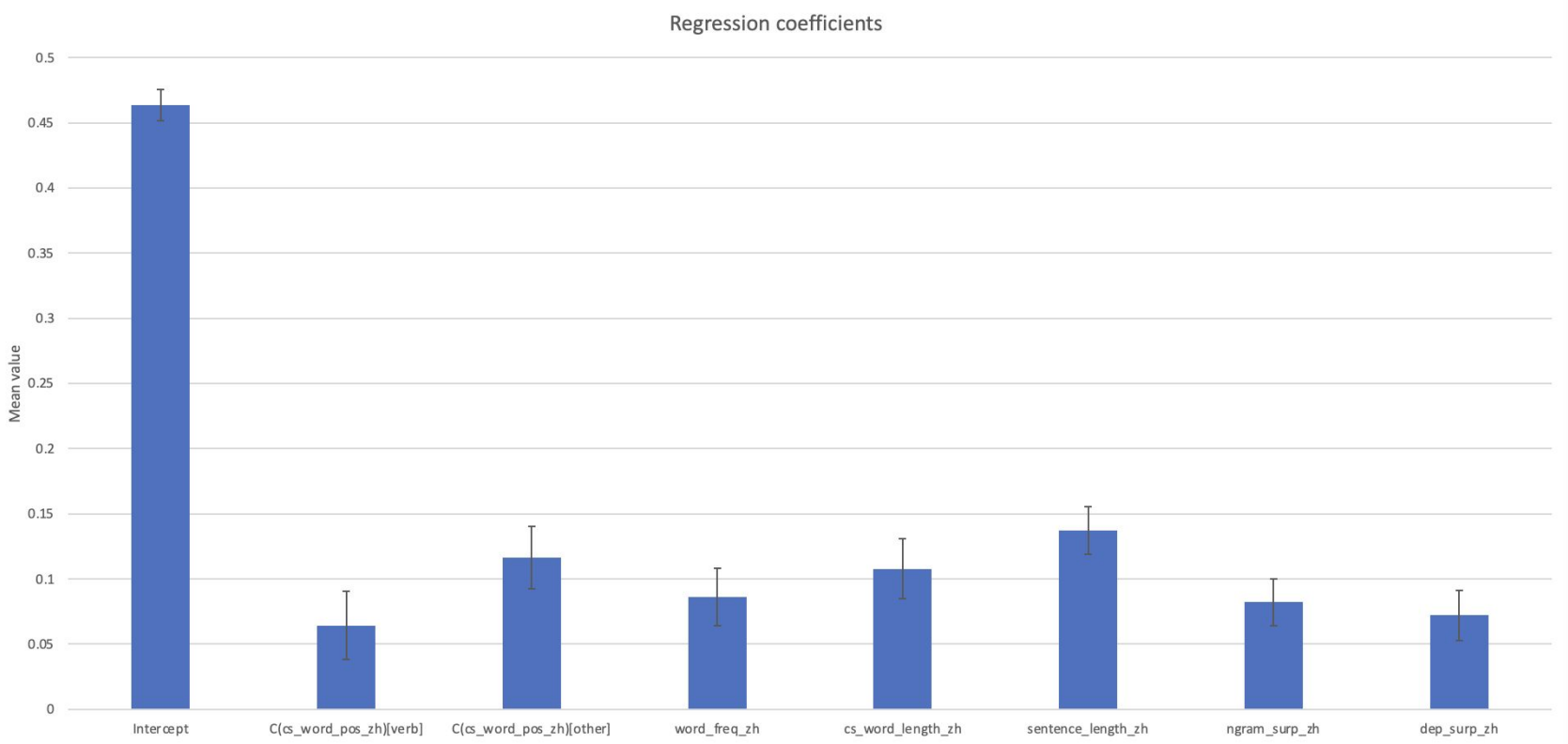


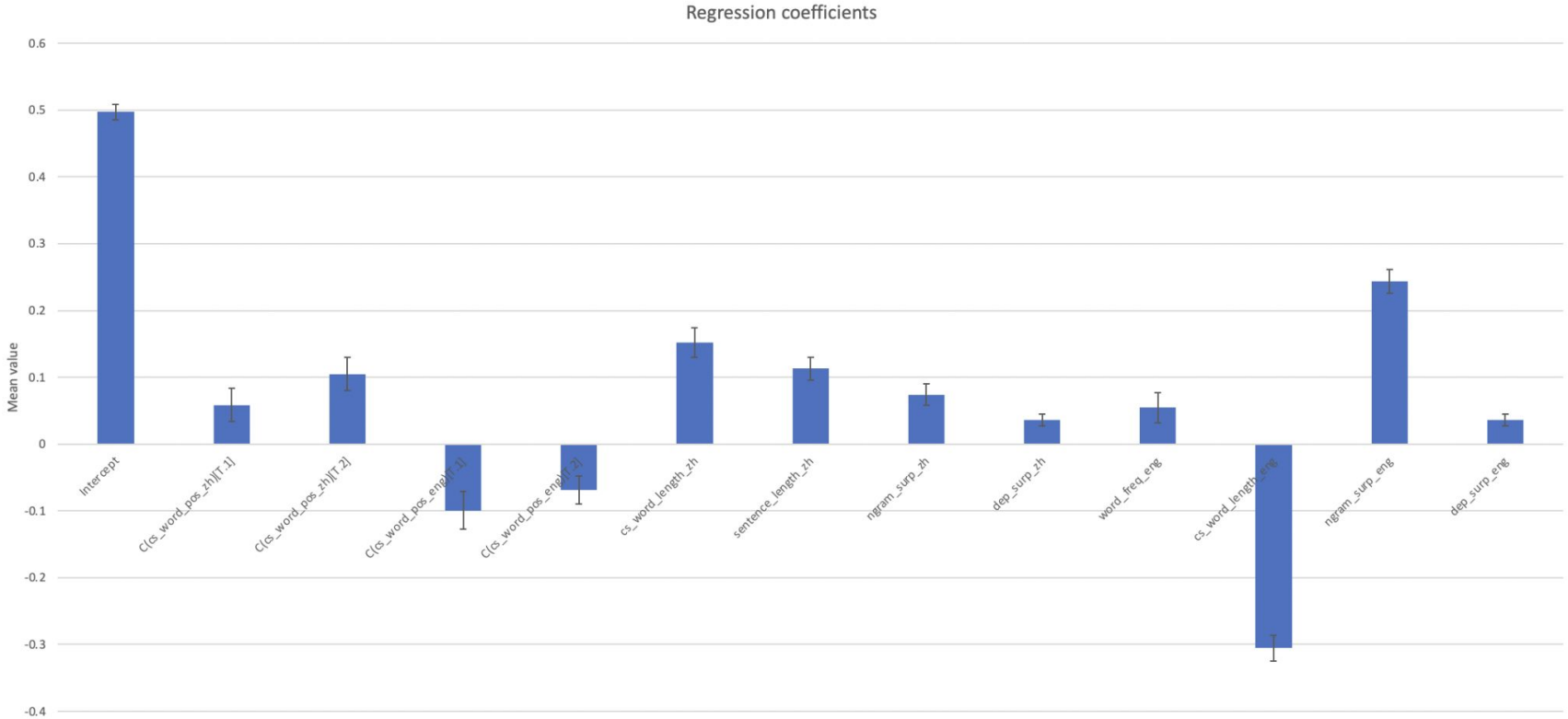
# Preliminary Results: comparing control to cs



# Preliminary Results: regression with zh features



# Preliminary Results: regression with zh and en features



# Upcoming plans

- Train better *n-gram* models

- Qualitative analysis
  - When does the model predict wrong?

- Classify code-switches into categories from Tsoukala et al. (2019)
  - Insertional switching
  - Alternational switching

# Tsoukala et al. code-switching categories

- Noun insertion:
  Target: un anfitrión feliz ha pateado un bolígrafo . (English: a happy host has kicked a pen)
  Output: un anfitrión feliz ha pateado un *pen* .
- Verb insertion:
  Target: un camarero llevó la llave . (English: a waiter carried the key)
  Output: un camarero *carried* la llave .
- Determiner insertion:
  Target: he is showing the book to the father .
  Output: he is showing *el* book to the father .
- Adjective insertion:
  Target: a man is sad . Output: a man is *triste* .

- Alternation at the determiner (Noun Phrase):
  Target: the uncle has shown a father the toy .
  Output: the uncle has shown *un padre* the toy .
- Alternation at the noun:
  Target: the short boy shows a brother a book .
  Output: the short boy shows a *libro a un hermano* .
- Alternation at the preposition (Prepositional Phrase):
  Target: the tall waiter has given a brother a book .
  Output: the tall waiter has given *a un hermano un libro* .
- Alternation at the auxiliary verb (Auxiliary Phrase):
  Target: the short waiter is showing a dog a toy .
  Output: the short waiter *está mostrando a un perro un juguete* .

# Thank you!