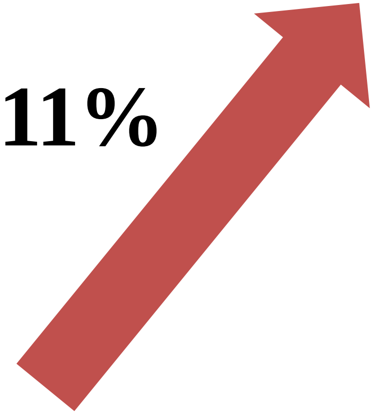
Chinese-English Code-Switching Analysis: The Case of Stack Exchange Posts

Qidu Fu

April 14, 2025

23% of conversations mix at least two languages among multilingual speakers (Quick et al., 2018).



+ 7%-11%

11%-16% of conversations mix at least two languages between a multilingual and mono-lingual (Montanari et al. 2019).

Agenda

Background: overview

Significance

Theoretical framework and research question

Analysis and finding

- Data
- Analysis

Agenda

Background: overview

Significance

Theoretical framework and research question

Analysis and finding

- Data
- Analysis

Overview: code-switching

Define code-switching (CS)

- CS refers to the switching from one language to another during the course of a conversation (Tulloch et al., 2023).

Know more about CS

- **23%** of conversations involves CS among multilingual speakers (Quick et al., 2018). CS is commonly seen in multilingual speakers' conversations.

Agenda

Background: overview

Significance

Theoretical framework and research question

Analysis and finding

- Data
- Analysis

Significance

Cultural dynamics

- Reveal cultural adaptation and linguistic dynamics between Chinese and English in online communities.

Educational insights

- Help educators and platform developers support bilingual communities.

Agenda

Background: overview

Significance

Theoretical framework and research question

Analysis and finding

- Data

- Analysis

Theoretical foundation

Theory: situational code-switching
CS occurs due to external factors such as setting, topic, or changes in the social situation (Bassiouney, 2020).

Example 1: professional VS personal
Switch from English for work-related discussions to a native language for personal matters.

Example 2: general VS cultural
Switch from native language to English when discussing general topics, but use the native language for cultural discussions.

Research question

RQ1: How frequently does code-switching occur in social media posts on a Stack Exchange site?

RQ2: Is there a significant difference in code-switching frequency across different discourse domains/topics?

RQ3: Does part-of-speech (POS) shifting occur in code-switching words?

Agenda

Background: overview

Significance

Theoretical framework and research question

Analysis and finding

- Data

- Analysis

Agenda

Background: overview

Significance

Theoretical framework and research question

Analysis and finding

- **Data**

- Analysis

Dataset

Stack Exchange API

Stack Exchange dataset 12,041 records/rows of data

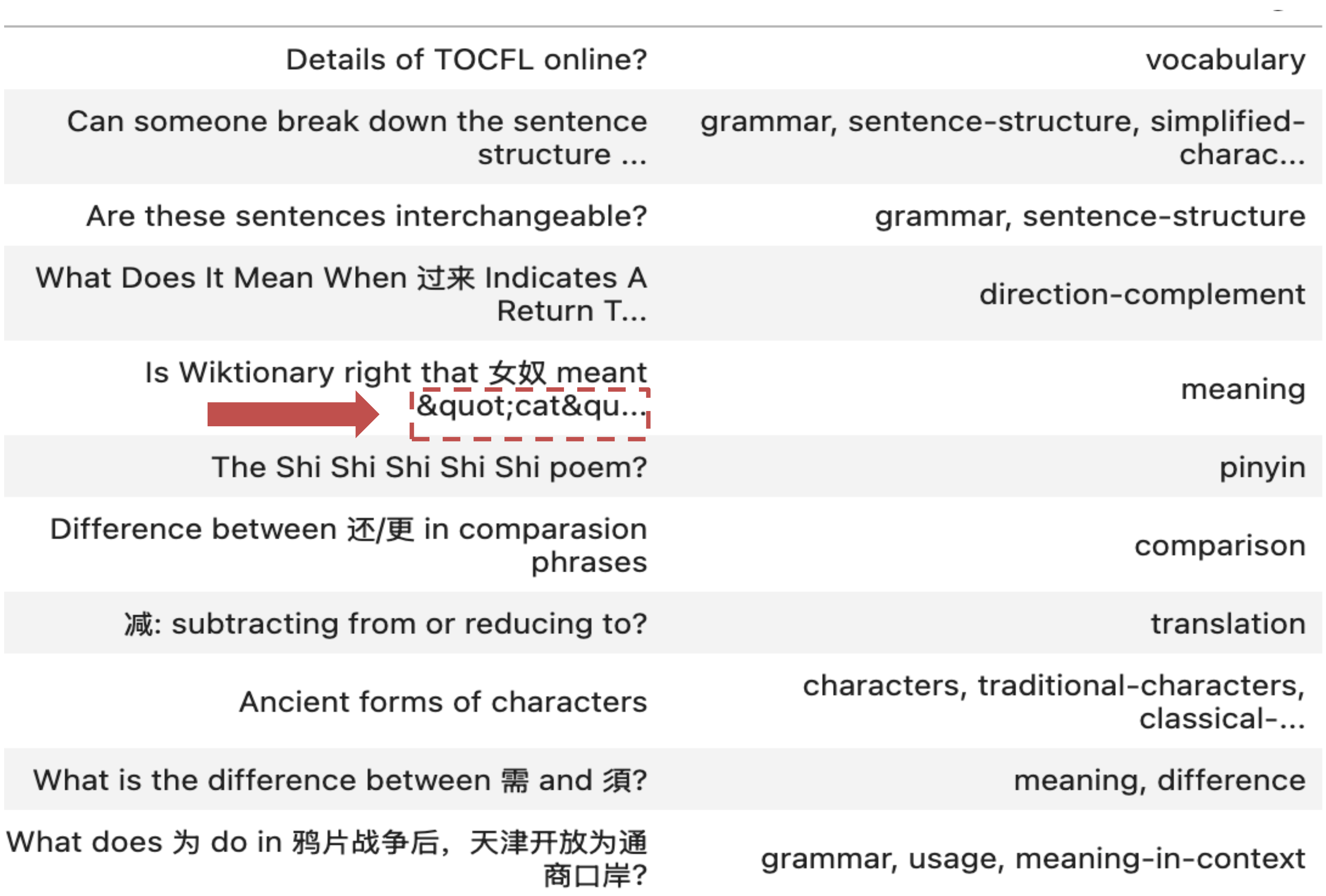
GitHub

SEAME dataset 5,321 rows

HuggingFace

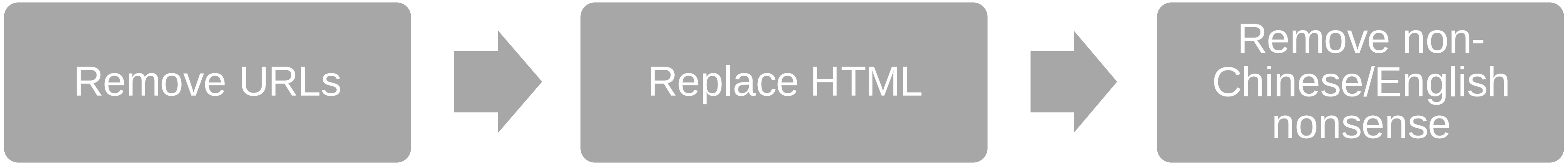
ASCEND dataset 9,869 rows

Data



Details of TOCFL online?	vocabulary
Can someone break down the sentence structure ...	grammar, sentence-structure, simplified-charac...
Are these sentences interchangeable?	grammar, sentence-structure
What Does It Mean When 过来 Indicates A Return T...	direction-complement
Is Wiktionary right that 女奴 meant &quot;cat&qu...	meaning
The Shi Shi Shi Shi Shi poem?	pinyin
Difference between 还/更 in comparasion phrases	comparison
減: subtracting from or reducing to?	translation
Ancient forms of characters	characters, traditional-characters, classical-...
What is the difference between 需 and 須?	meaning, difference
What does 为 do in 鸦片战争后，天津开放为通商口岸?	grammar, usage, meaning-in-context

Data cleaning



Remove URLs

Replace HTML

Remove non-Chinese/English nonsense

Agenda

Background: overview

Significance

Theoretical framework and research question

Analysis and finding

- Data

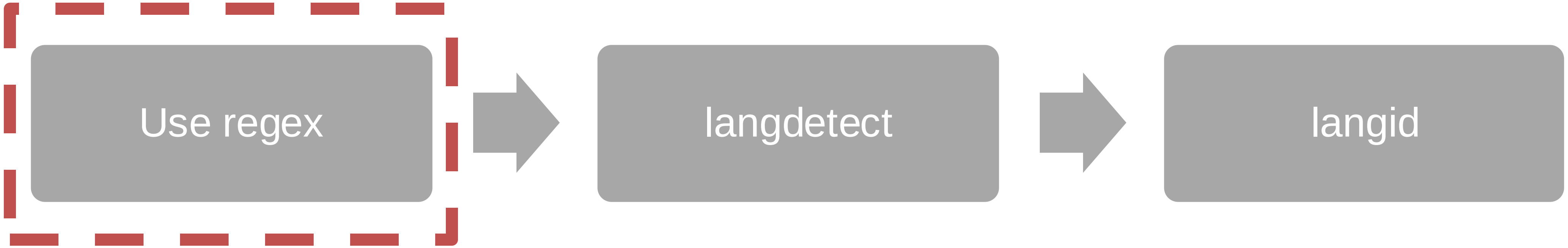
- **Analysis**

Data tokenization

Stanza: English and Chinese models

- **12,041** records/rows
- **152,573** word tokens
- **13,664** sentences
- a mean word count of ~**10** per sentence

Modeling: CS detection



Use regex

langdetect

langid

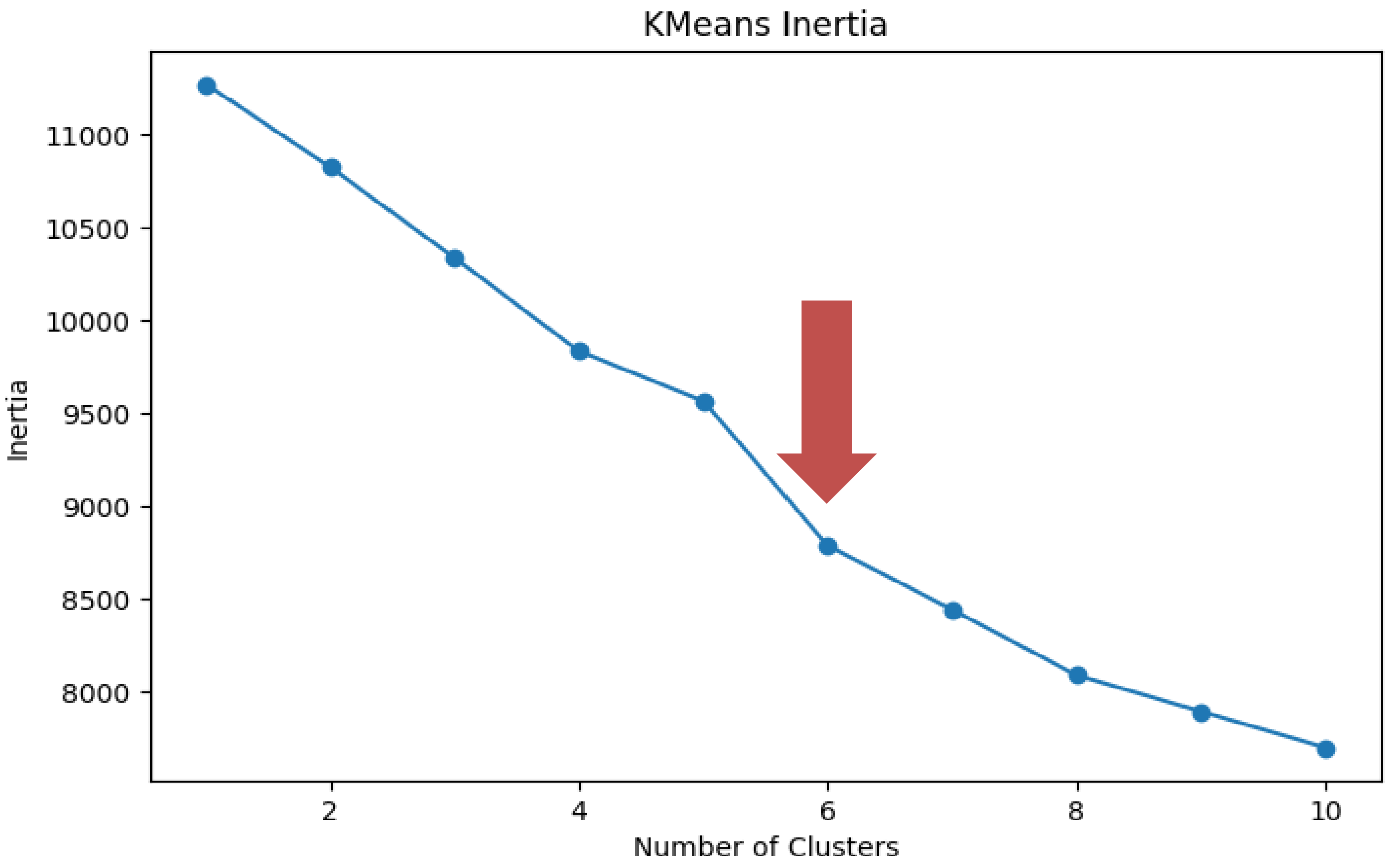
Modeling: topic modeling

K-Means

Use LDA

K-Means

Identify the number of topics: **6**

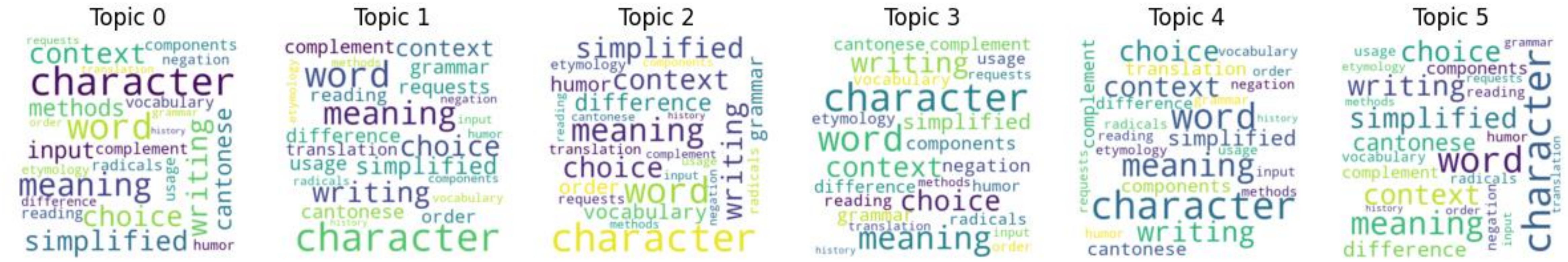


Modeling: topic modeling

K-Means

Use LDA

LDA: top words in WordCloud



Modeling: topic modeling

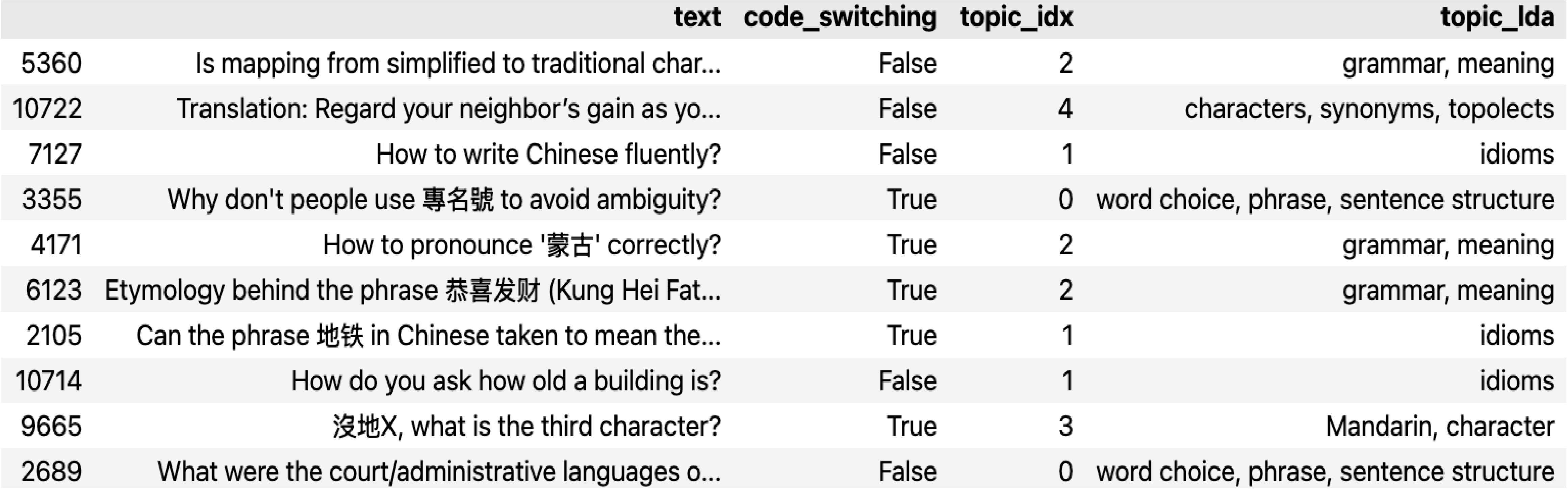
K-Means

Use LDA

Manually put forward descriptive names for each topic

- Top words
- STDs

CS distribution across domains: RQ1 & RQ2

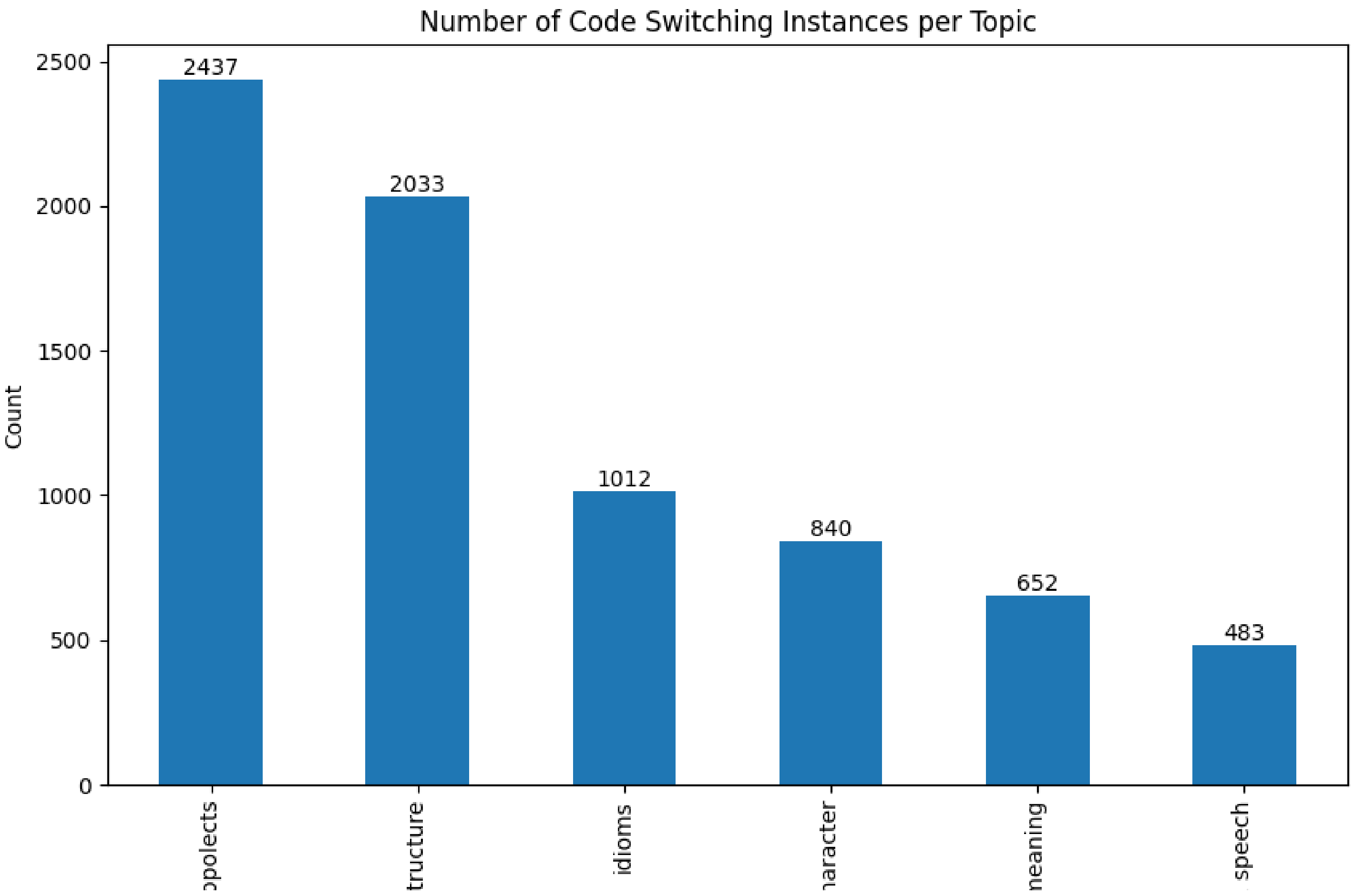


	text	code_switching	topic_idx	topic_lda
5360	Is mapping from simplified to traditional char...	False	2	grammar, meaning
10722	Translation: Regard your neighbor’s gain as yo...	False	4	characters, synonyms, topolects
7127	How to write Chinese fluently?	False	1	idioms
3355	Why don't people use 專名號 to avoid ambiguity?	True	0	word choice, phrase, sentence structure
4171	How to pronounce '蒙古' correctly?	True	2	grammar, meaning
6123	Etymology behind the phrase 恭喜发财 (Kung Hei Fat...	True	2	grammar, meaning
2105	Can the phrase 地铁 in Chinese taken to mean the...	True	1	idioms
10714	How do you ask how old a building is?	False	1	idioms
9665	没地X, what is the third character?	True	3	Mandarin, character
2689	What were the court/administrative languages o...	False	0	word choice, phrase, sentence structure

CS distribution across domains: RQ1 & RQ2

Topic and CS counts (*descending order*)

- 4 characters, synonyms, topolects **2,437**
- 0 word choice, phrase, sentence structure **2,033**
- 1 idioms **1,012**
- 3 Mandarin, character **840**
- 2 grammar, meaning **652**
- 5 difference: writing, speech **483**

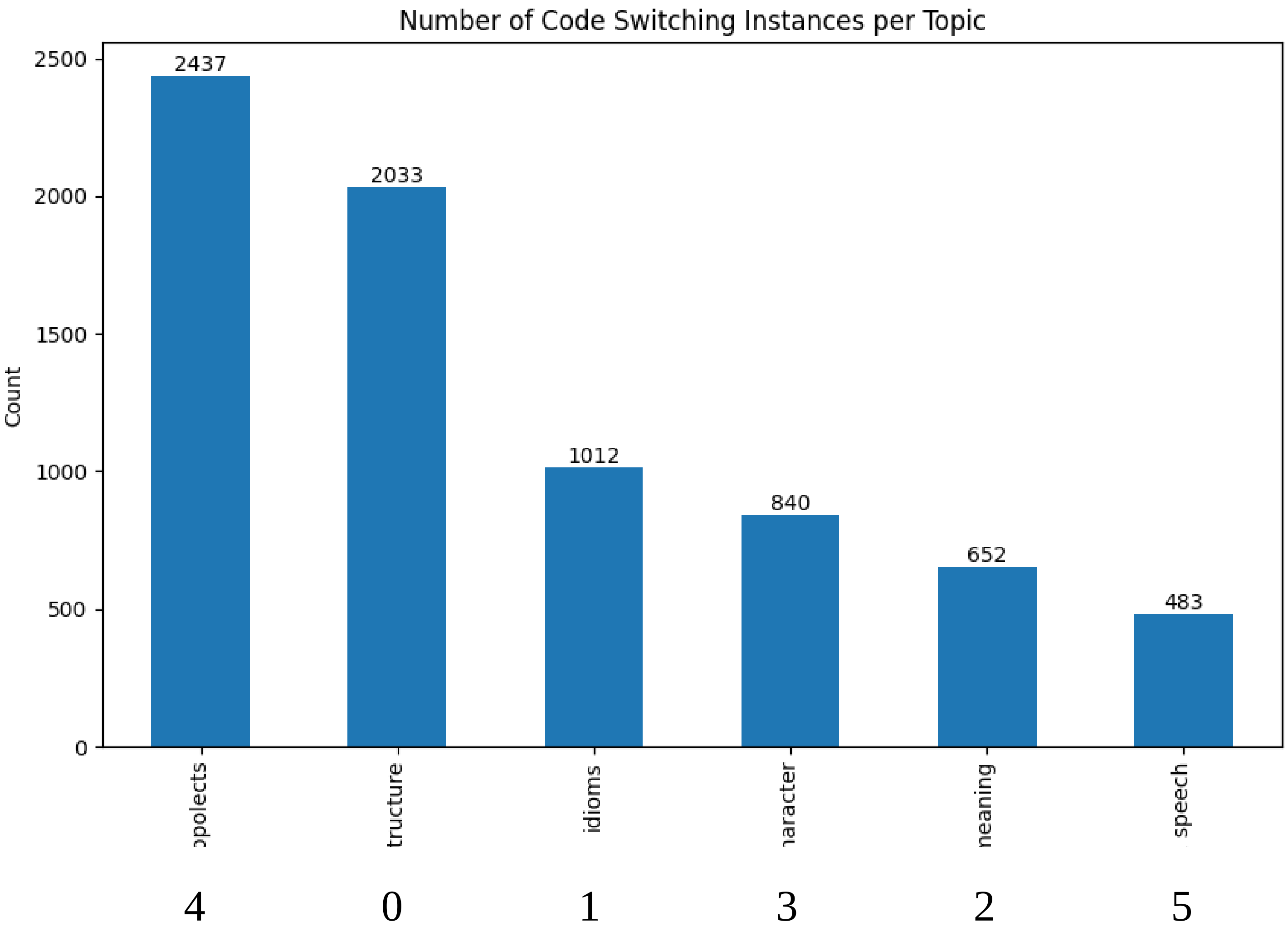


4 0 1 3 2 5

CS distribution across domains: RQ1 & RQ2

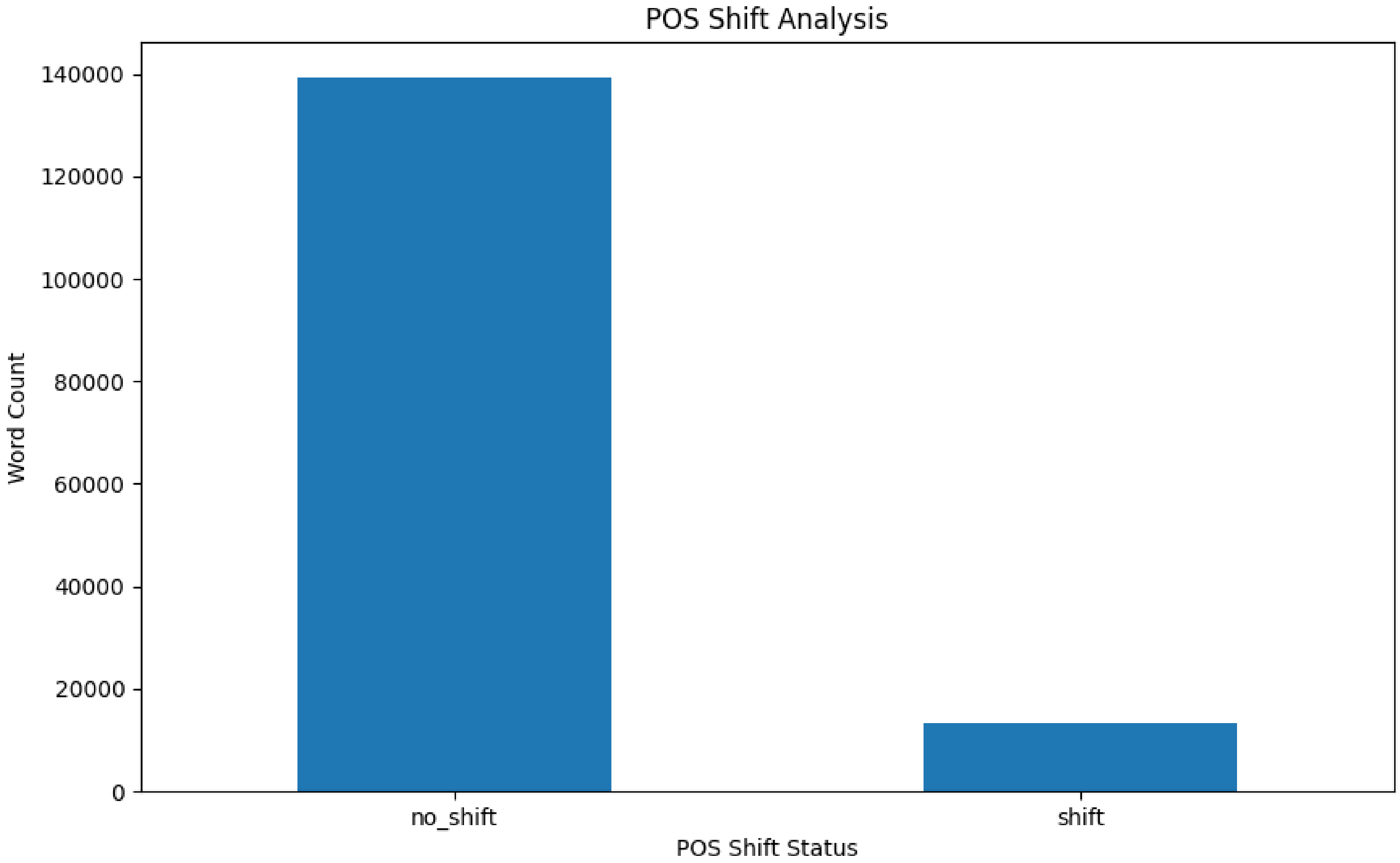
Total of CS: **7,457**

Kruskal_wallis test: p = **0.415**



POS shifting in CS: RQ3

- Non-shift **139,302**
- shift **13,271**



POS shifting in CS: RQ3

Shifting examples

- Need help to translate 陈子昂's 《感遇·翡翠巢南海》

  V N (personal’s name)

- "Usage of ""我"" in Li Bai's lines ""东风随春归，发我枝上花""”

  Prep Pro (I/me)

Non-shifting examples

- "Can the term ""老大"" be used to call ""Boss"" in video games nowadays?”

  N N (boss)

- "Usage of ""我"" in Li Bai's lines ""东风随春归，发我枝上花""”

  N N (poetry lines)

Thank you!

Questions?

Presenter: Qidu Fu Email:
qiduf@andrew.cmu.edu

References

- Bassiouney, R. (2020). *Arabic sociolinguistics: Topics in diglossia, gender, identity, and politics*. Georgetown University Press.

- Montanari S, Ochoa W, & Subrahmanyam K. (2019). A longitudinal investigation of language mixing in Spanish–English dual language learners: The role of language proficiency, variability, and sociolinguistic factors. *Journal of Child Language*, 1, 1–25. 10.1017/S0305000919000278

- Quick, A. E., Lieven, E., Carpenter, M., & Tomasello, M. (2018). Identifying partially schematic units in the code-mixing of an English and German speaking child. *Linguistic Approaches to Bilingualism*, *8*(4), 477–501. https://doi.org/10.1075/lab.15049.qui

- Tulloch, M. K., & Hoff, E. (2023). Filling lexical gaps and more: Code-switching for the power of expression by young bilinguals. *Journal of Child Language*, *50*(4), 981–1004. https://doi.org/10.1017/S0305000922000307