Examining the Replicability of Grammaticality Judgments in Chinese Journal Articles: Dialectal Influences and Sources of Variability

Hai Hu[1], Aini Li[2], Yina Patterson[3], Jiahui Huang[4]
Chien-Jer Charles Lin[5]

1: Shanghai Jiao Tong University
2: University of Pennsylvania
3: Brigham Young University
4: University of Washington
5: Indiana University Bloomington

2022 March, HSP@UCSC

Outline

1. Background
2. Research questions
3. Methods and data
4. Experiment 1: acceptability rating
5. Experiment 2+3: forced-choice task
6. Discussion
7. Conclusion

Background

- Grammaticality judgments are central to linguistic research

  a. [Who]$_i$ did he claim [that he met t_i] ?

  b. *[Who]$_i$ did he make [$_{\text{NP complex island}}$ the claim [that he has met t_i]] ?

- Doubts about whether informal judgments are reliable (Gibson et al 2010, 2013a, 2013b)

- Different ways to think about grammaticality judgments (gradient vs. binary) (Francis, 2022)

Background: replicating informal judgments

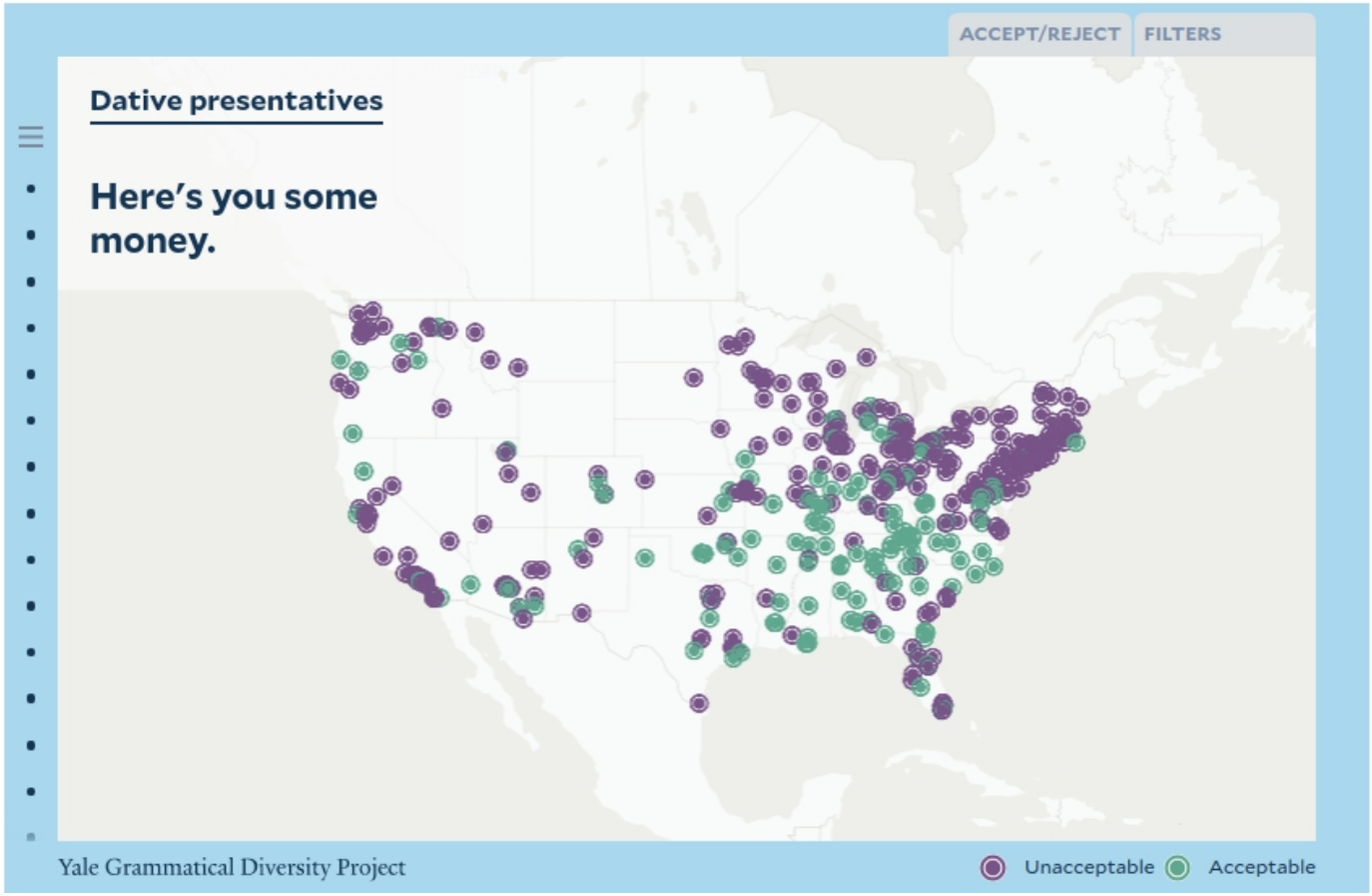
- Informal judgments =?= judgments under experimental setting

Language	Sources of stimuli	Convergence rate	
English	Syntax textbook *Core Syntax* (Adger 2003)	**Likert Scale**: 97.4% **Forced Choice**: 98%	Sprouse&Almeida '12
English	Journal: *Linguistic Inquiry*	**Likert Scale**: 95% **Forced Choice**: 95%	Sprouse et al '13
Japanese and Hebrew	Journal articles: 'Potentially questionable' examples	**Likert Scale**: Hebrew: 50% Japanese: 71.43%	Linzen&Oseki '18
Chinese	Syntax textbook: *The Syntax of Chinese* (Huang et al 2009)	**Likert Scale**: 89.2% **Forced Choice**: 96.8%*	Chen et al '20

12 out of 17 problematic pairs; 153/158 pairs=96.8%

Background: Dialectal influence on grammatical diversity

- Yale Grammatical Diversity Project (Zanuttini et al 2018)



Research questions

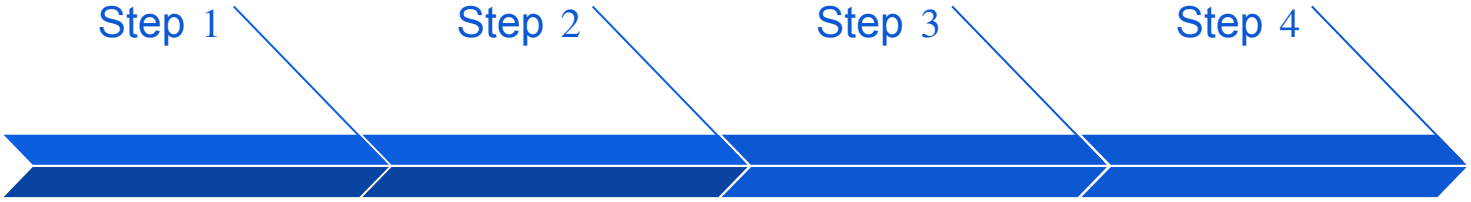
Gap 1: for non-English languages, a more representative sample

→ RQ1: How reliable are the informal judgments for Chinese sentences from a wide range of journal articles, compared with ones obtained under stricter experimental setting?

Gap 2: other factors: participants' backgrnd, author backgrnd

→ RQ2: What other factors influence judgments, e.g., **dialectal/language background** of participants/authors, age, gender, etc.

Method: obtain stimuli



Step 1	Step 2	Step 3	Step 4
10 journals	**Sample 68 articles on Chinese syntax**	**Sample 6 ungram (* or ??) sentences per paper**	**337 minimal pairs**
2010-2020 Published in Chinese (2) English (7) both (1)	7261 example sentences	Find/construct minimal pairs Remove: anaphora, interpretation, prosody	Stimuli for Exp1

Method: participant background

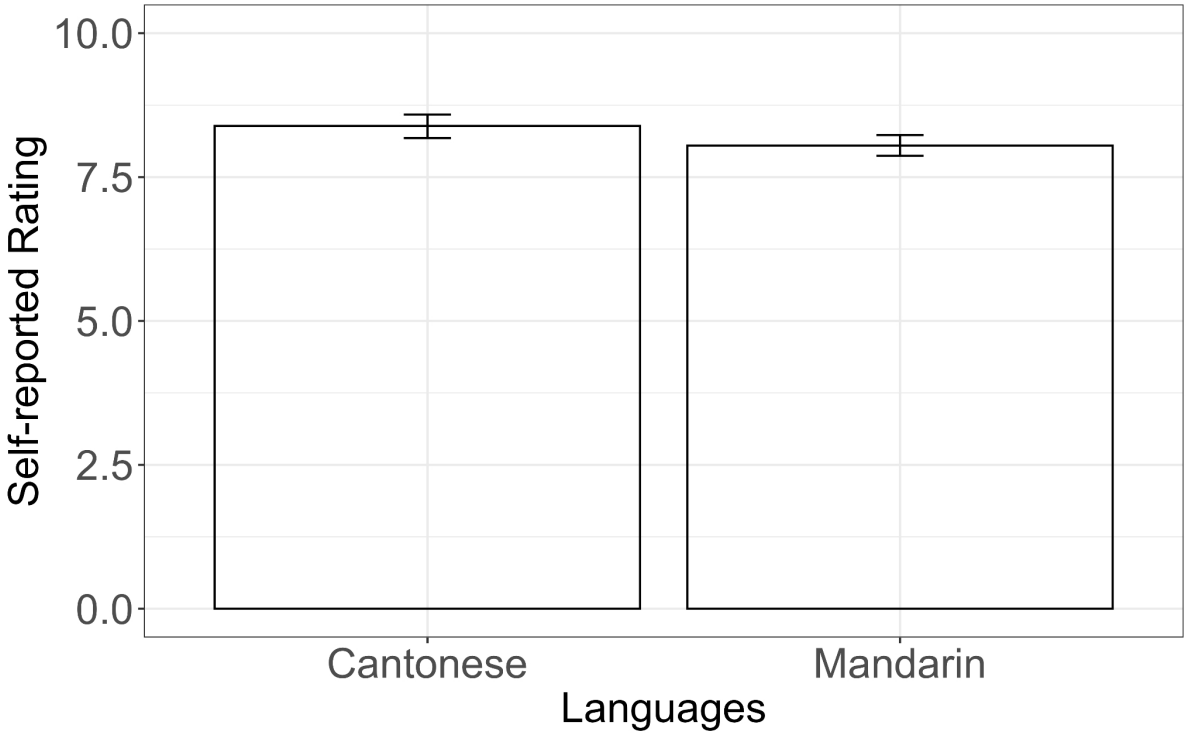
Two dialect/language background (**regions**):

1. Beijing (**BJ**): native speakers of Mandarin (N of monolinguals = 161/187)

Method: participant background

Two dialect/language background (**regions**):

1. Beijing (**BJ**): native speakers of Mandarin (N of monolinguals = 161/187)
2. Guangzhou (**GZ**): bilingual speakers of Mandarin and Cantonese



Method: Mandarin vs. Cantonese

- Almost mutually incomprehensible:
  - Sound differences are drastic (Tang and van Heuven 2009)

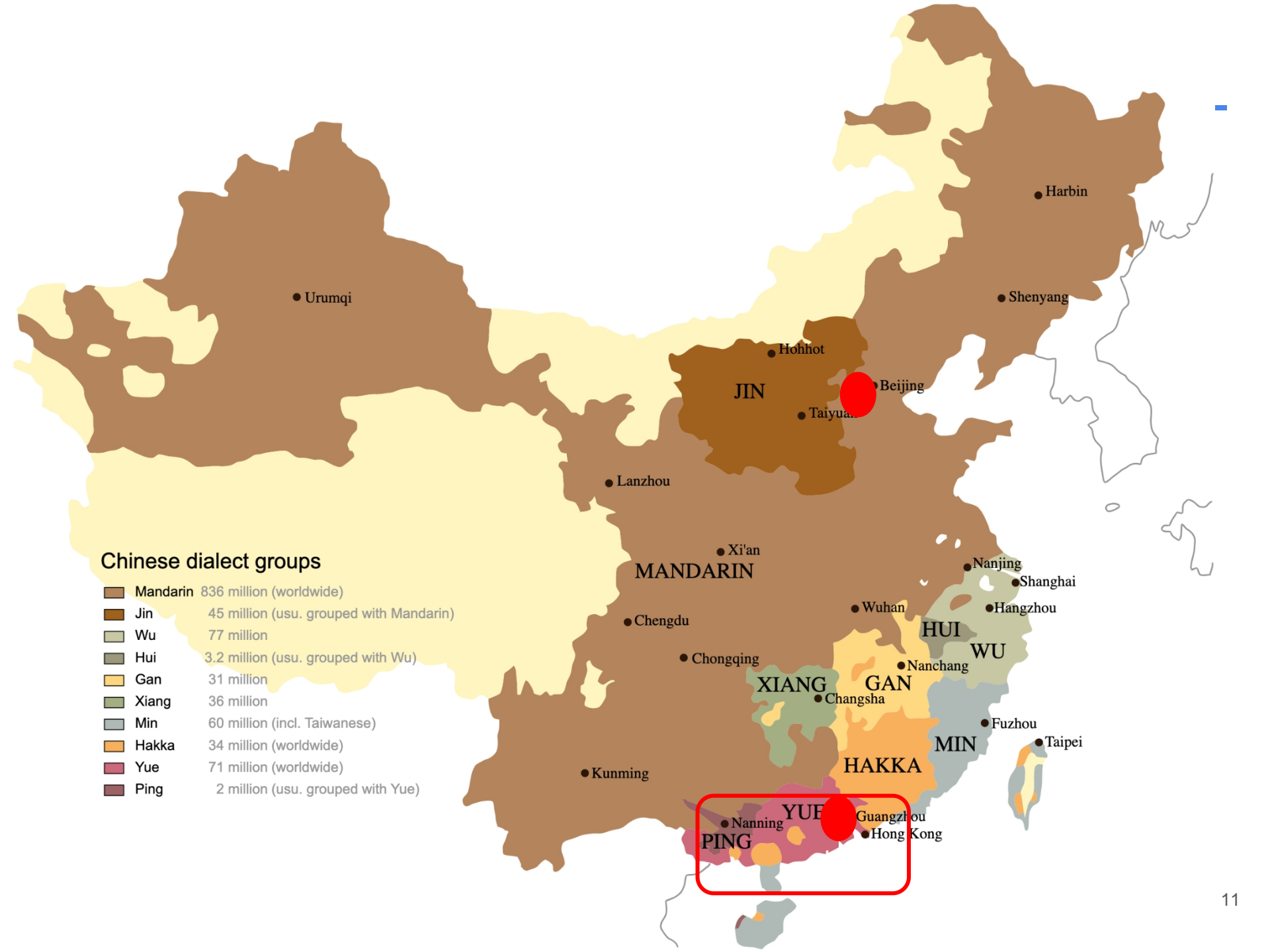
63%
Guangzhou Cantonese —--> Beijing Mandarin
<—--
34%

  - Lexical differences exist along with shared cognates

Cantonese newspapers unintelligible to Mandarin speakers, more easily vice versa (Zhang 1998)

  - Differences in syntax eg.,

Mandarin	Cantonese
VP -> ADV + V	VP -> V + ADV
VP -> V + not + complement	VP -> not + V + complement


Chinese dialect groups
Mandarin 836 million (worldwide)
Jin 45 million (usu. grouped with Mandarin)
Wu 77 million
Hui 3.2 million (usu. grouped with Wu)
Gan 31 million
Xiang 36 million
Min 60 million (incl. Taiwanese)
Hakka 34 million (worldwide)
Yue 71 million (worldwide)
Ping 2 million (usu. grouped with Yue)
Urumqi
Harbin
Shenyang
Hohhot
JIN
Beijing
Taiyuan
Lanzhou
Xi'an
MANDARIN
Nanjing
Shanghai
Wuhan
Hangzhou
Chengdu
HUI
WU
Chongqing
Nanchang
XIANG
GAN
Changsha
Fuzhou
MIN
Taipei
HAKKA
Kunming
YUE
Guangzhou
Nanning
Hong Kong
PING

Method: author background

- Coded the background of first author:

- 4 levels: mainland, Taiwan, Hong Kong, Other
- recoded later as mainland vs. non-mainland

- Operationalization:

- To the best of our/Internet's knowledge, where is the author before the age of 18?

Method: other factors

- Sentence length:
  - n characters (mean=10.22, std=4.56)
- Paper language:
  - Chinese (n=22) or English (n=46)
- Participants:
  - age, gender, education

Method: three experiments

Exp1: 337 pairs for 7-point Likert Scale judgment
How natural is the following sentence?

Method: three experiments

Exp 2 & 3: unreplicated pairs for **forced-choice** task
Which one is more natural?

Forced choice task is more sensitive to grammaticality

Method: when is a judgment ‘replicated’?

7 point Likert-Scale (Exp 1):

For each pair in each region, replicated:
If and only if:
rating(gram) > rating(ungram) **and**
t.test(rating(gram), rating(ungram)) < 0.05

Forced Choice (Exp 2 + 3):

For each pair in each region, replicated:
If and only if:
num(gram) significantly > num(ungram)

Experimental details

Online questionnaire distributed using **Qualtrics**

Exp 1: each sentence rated by roughly 30 participants

BJ: n=187, 142 female, mean age=22

GZ: n=191, 149 female, mean age=25

Exp 2: each pair rated by roughly 40 participants

BJ: n=40, 32 female, mean age=20

GZ: n=38, 36 female, mean age=20

Exp 3: each pair rated by roughly 40 participants

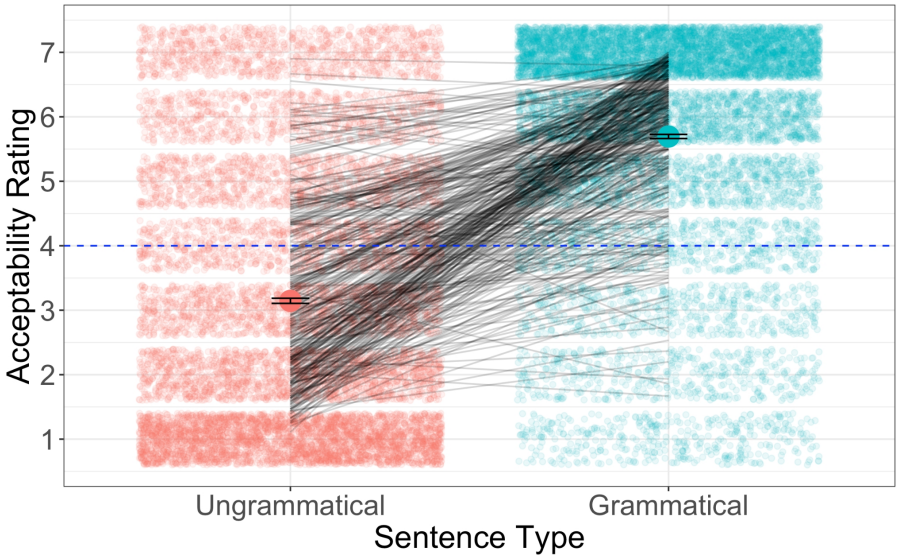
BJ: n=37, 31 female, mean age=22

GZ: n=49, 39 female, mean age=22

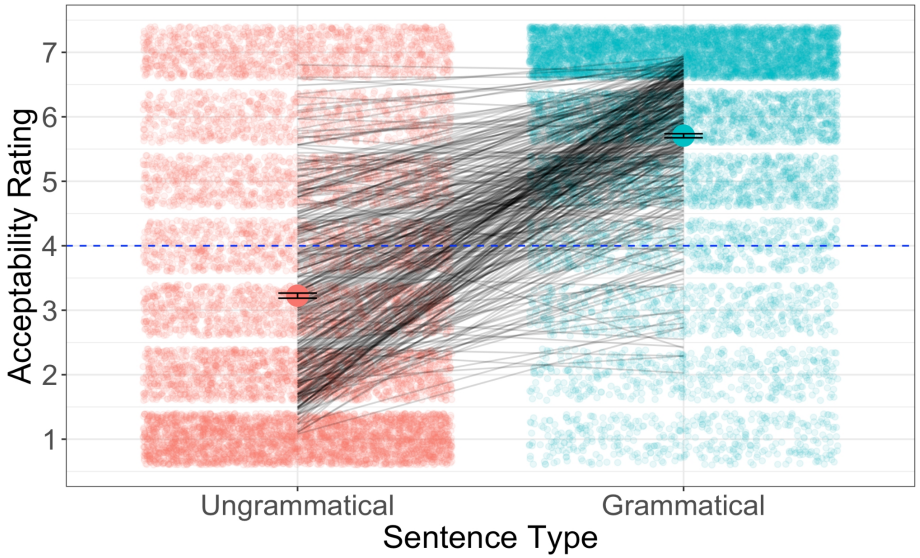
Exp 1 Results: mean rating

- Mean acceptability rating (raw scores)
- Beijing: Grammatical: 5.69 vs. Ungrammatical: 3.14
- Guangzhou: Grammatical: 5.71 vs. Ungrammatical: 3.23

Beijing participants



Guangzhou participants



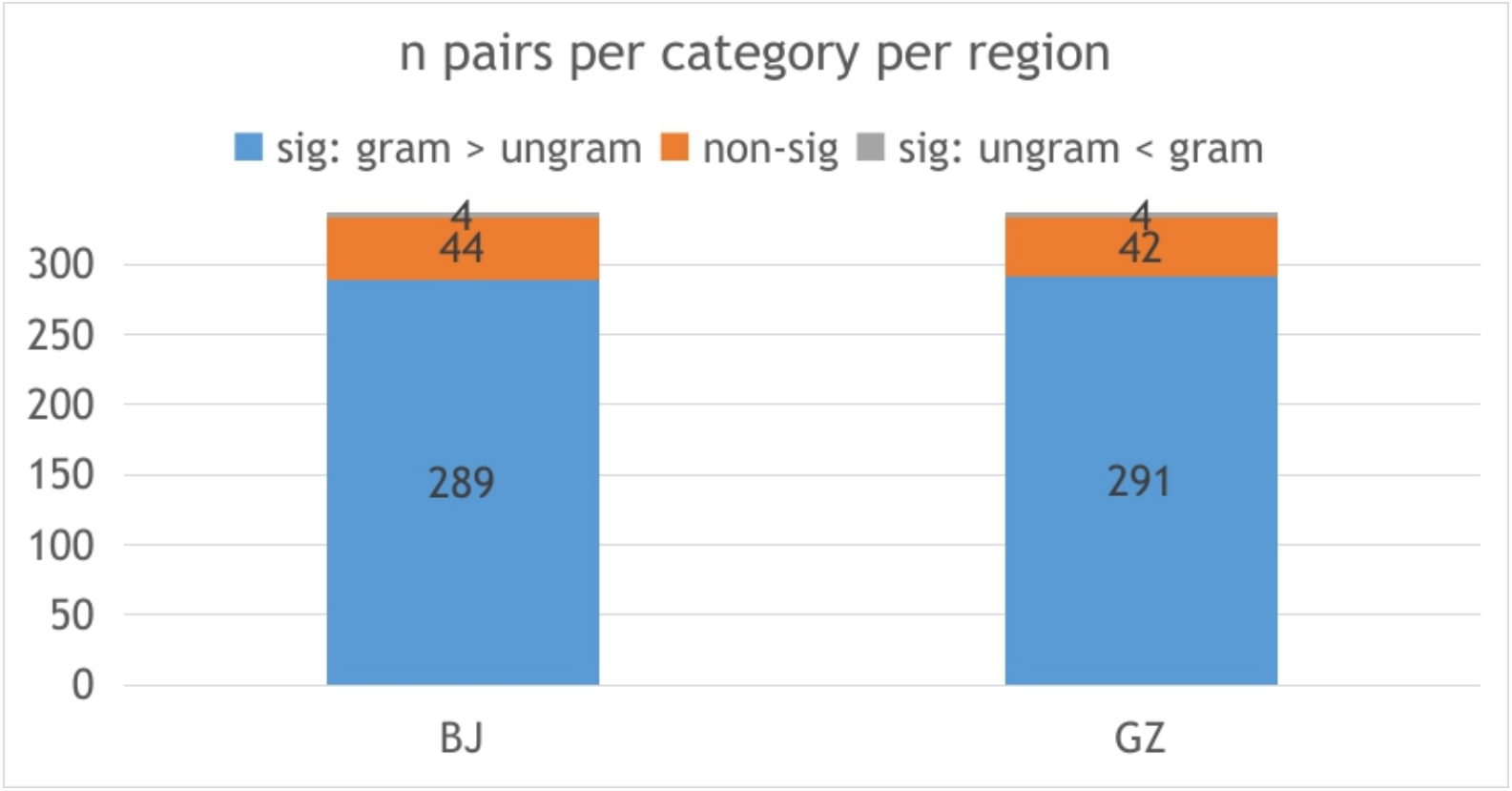
Exp 1 Results: regression model

Table 4: Modeling acceptability judgments: The results of liner mixed-effects regression

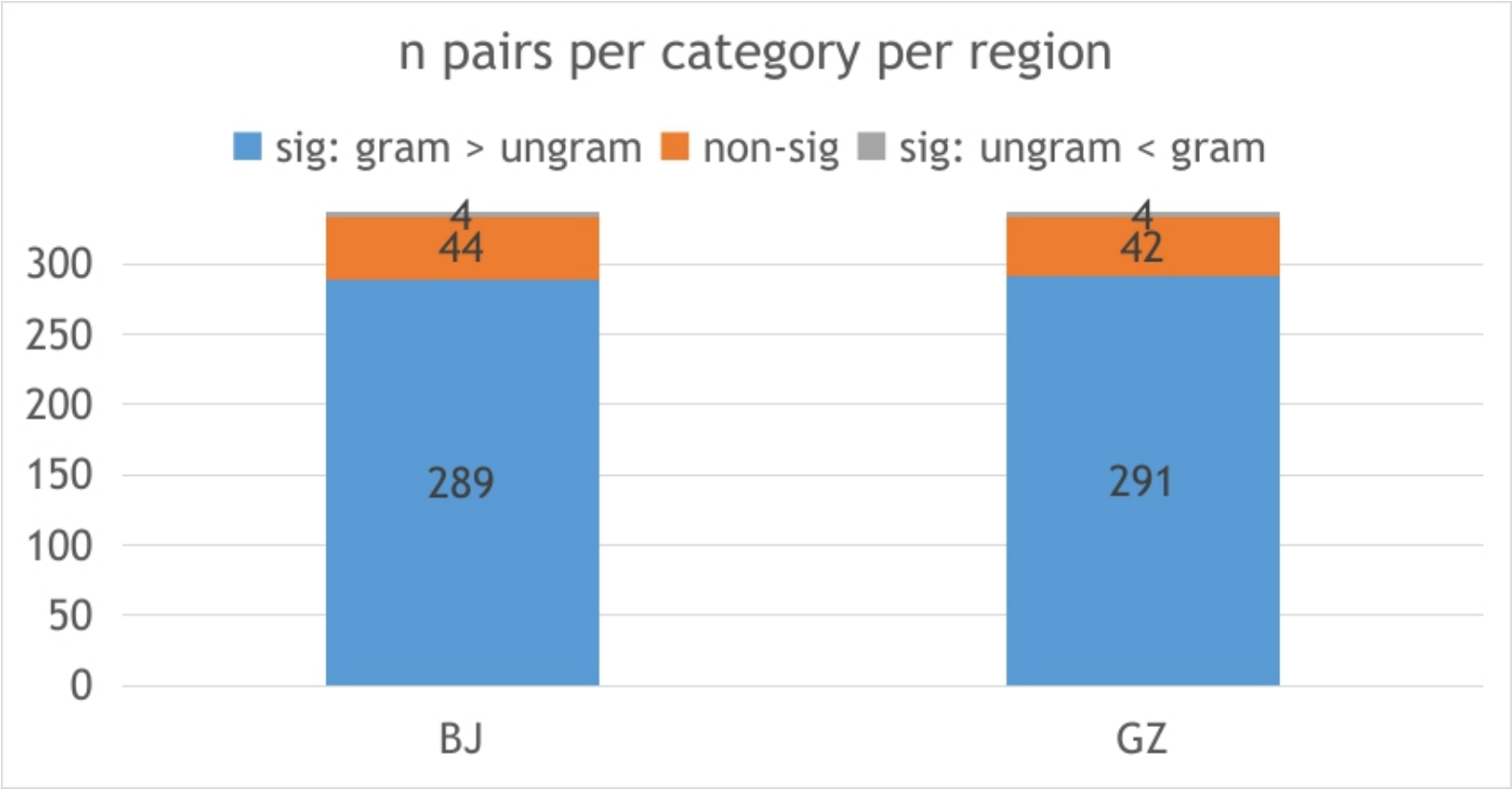
	Estimate	Std. Error	df	t value	Pr(>|t|)	
Intercept	3.734e-02	2.387e-02	9.063e+02	1.564	0.1181	
Grammaticality (Gram.)	1.129e+00	4.019e-02	6.428e+02	28.090	<0.001	***
Region (Beijing)	4.773e-05	6.408e-03	4.180e+04	0.007	0.99	
Education (BelowUndergrad)	9.864e-04	4.239e-02	4.181e+04	0.023	0.98	
Education (Undergrad)	9.215e-04	2.819e-02	4.184e+04	0.033	0.97	
Education (Master)	3.019e-06	3.028e-02	4.180e+04	0.000	0.99	
Age	2.703e-05	5.985e-04	4.187e+04	0.045	0.96	
Gender (Female)	-4.890e-04	7.562e-03	4.188e+04	-0.065	0.95	
First author's region (Mainland)	-4.752e-02	4.573e-02	6.422e+02	-1.039	0.30	
Paper language (English)	-2.028e-01	4.966e-02	6.426e+02	-4.085	<0.001	***
Sentence length	-4.085e-02	1.983e-02	6.385e+02	-2.060	0.04	*
Grammaticality (Gram.) : Region (Beijing)	1.512e-02	1.226e-02	4.179e+04	1.234	0.22	

- Grammatical sentences were rated higher
- Region: NOT significant
- First author's region: NOT significant
- Sentences in papers written in English rated lower
- Longer sentences rated lower (c.f. Yao et al 2018)

Exp 1 Results: convergence rate



Exp 1 Results: convergence rate



Convergence rate:

BJ: 289/337 pairs = 85.8%

GZ: 291/337 pairs = 86.4%

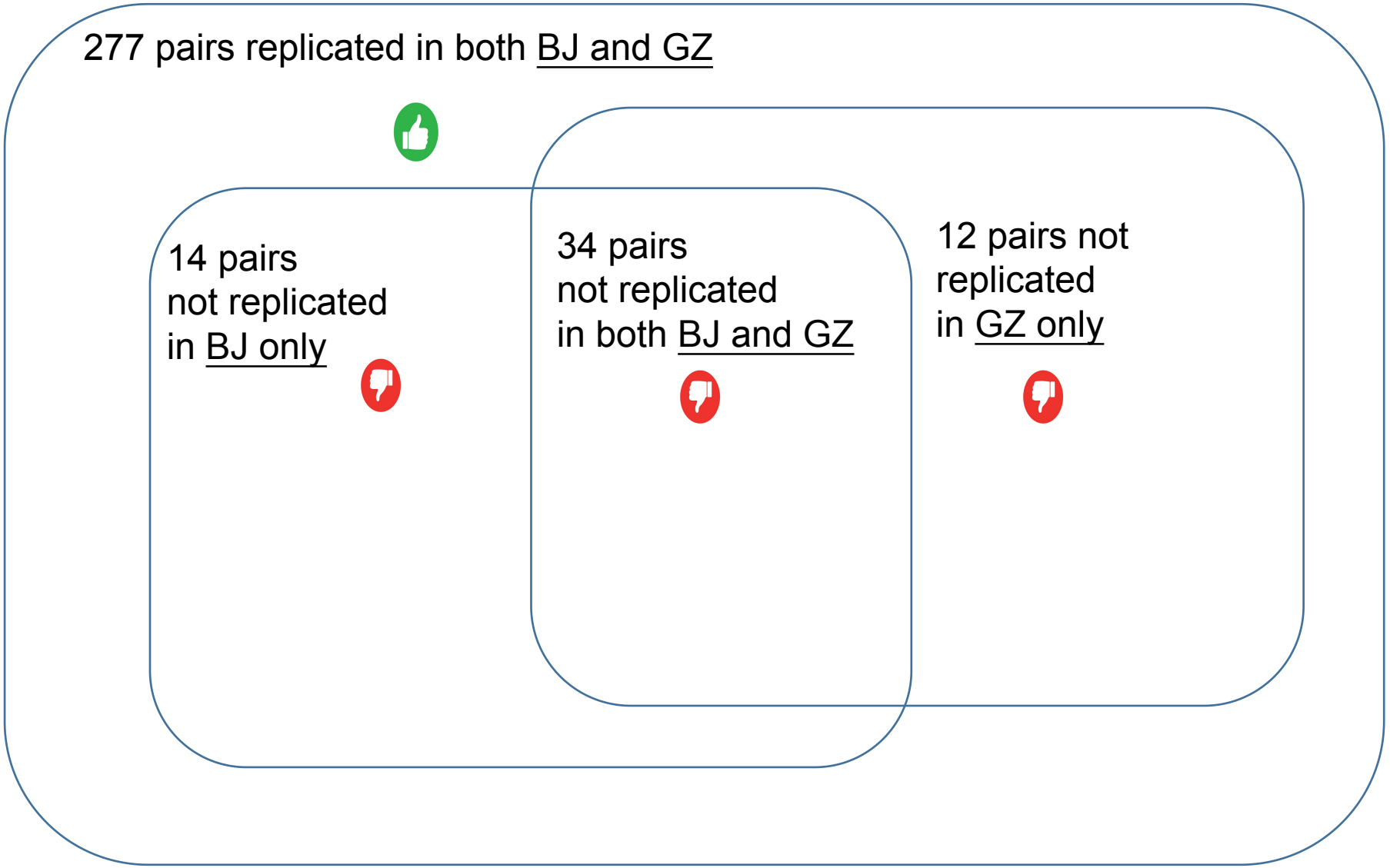
cf. English sentences in *Linguistic Inquiry*: ~95% (Sprouse et al 2013)

cf. Chinese sentences in textbook: 89.2% (Chen et al 2020)

Exp 2 + 3 Results

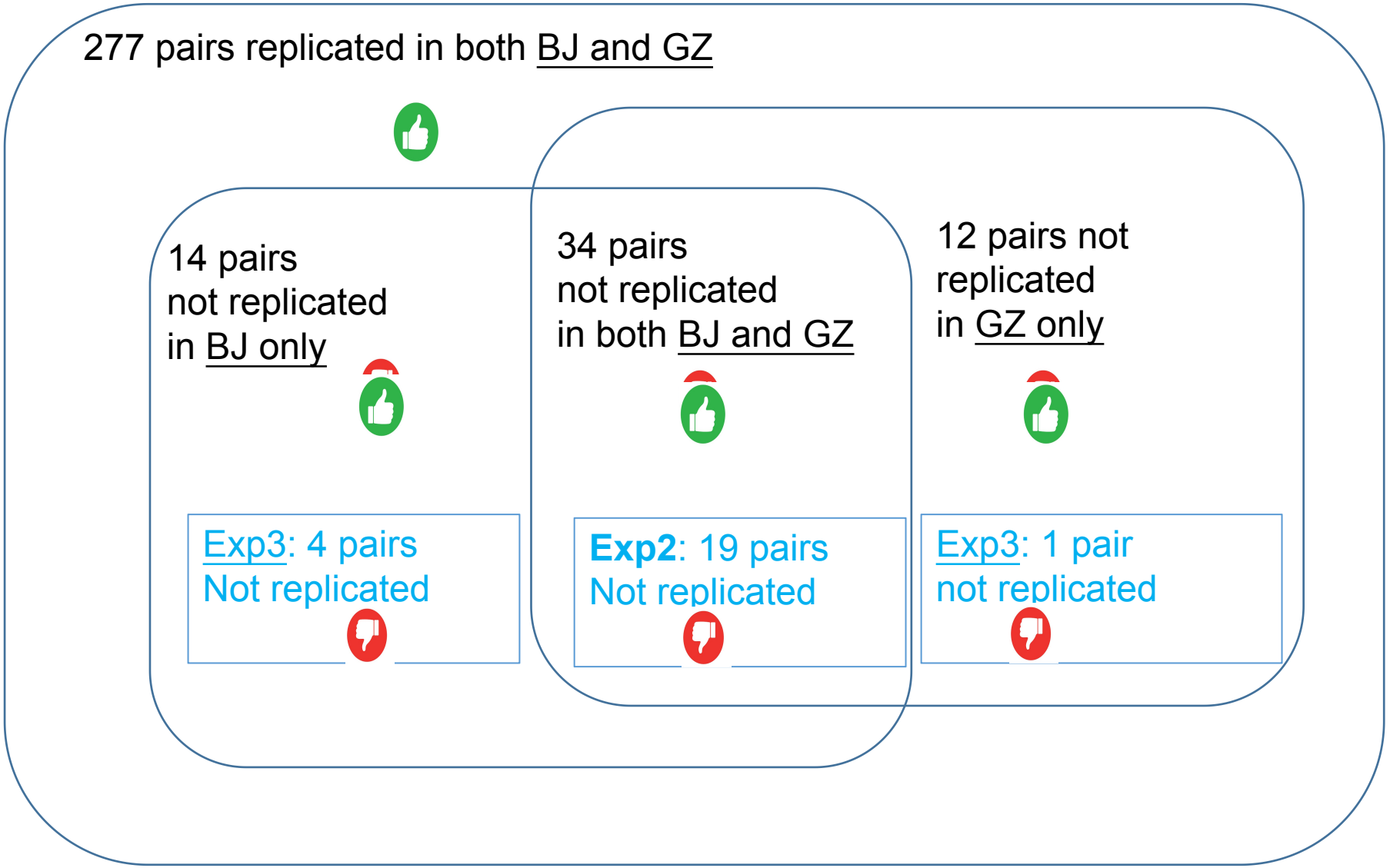
Exp 1: Likert Scale

Exp 2 + 3 Results



Exp 2 + 3 Results

Exp 1: Likert Scale
Exp 2 + 3: Forced Choice



Exp 2 (forced choice) Results

BJ and GZ have exactly the same pattern.

Categorization of these **19** unreplicated cases:

problematic	N =11; 58% (3% of 337)	Ex. NPIs, adversity passive voice, topic & focus
Semantic/ pragmatic	N = 6; 32% (2% of 337)	Ex. sentences need more discourse
other	N = 2; 11% (1% of 337)	Ex. one sentence from footnote

Exp 2 (forced choice) Results

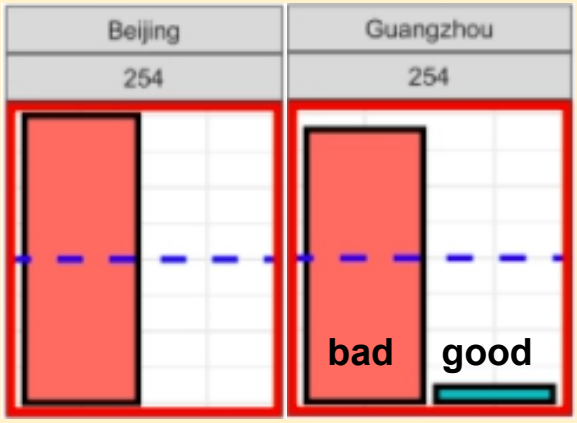
Examples of problematic cases:

fang2zhi3 (implicit negative verb) -> cong2lai2 NPI

a. 中国　古代　从来（*没有）防止　人口　流动

China ancient time **NPI** **(*no)** **prevent** population flow

'Ancient China has always prevented population flow.'



2014: 579)

Exp 2 (forced choice) Results

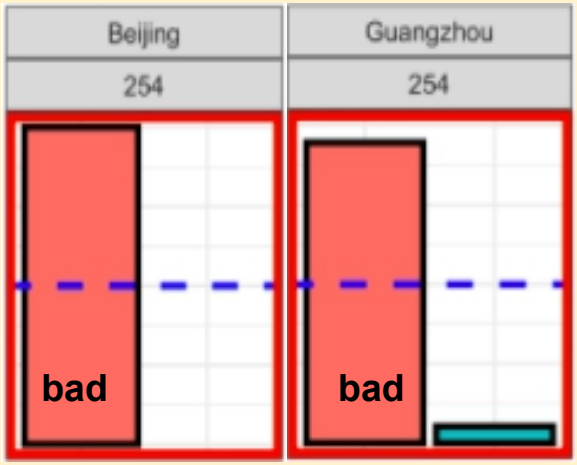
Examples of problematic cases:

fang2zhi3 (implicit negative verb) -> cong2lai2 NPI

a. 中国　古代　从来（*没有）防止　人口　流动

China ancient time **NPI** **(*no)** **prevent** population flow

'Ancient China has always prevented population flow.'



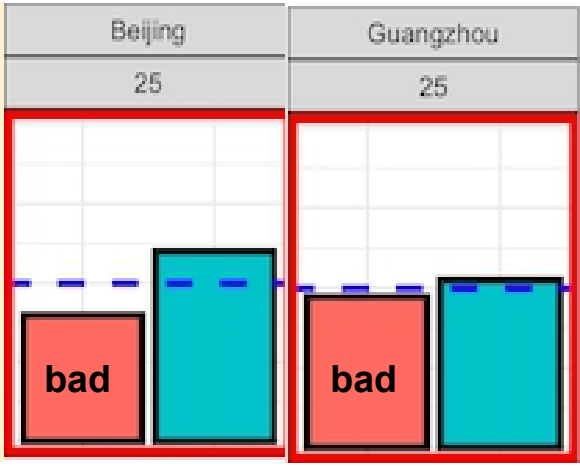
2014: 579)

adversity *BEI* passive voice -> undesirable verbs

b. 我　被　批评/*表扬　了

I **BEI** **criticize** / ***praise** LE

"I was criticized/praised."　(Liu 2011: 215, cited from Li & Thomson, 1989)



Exp 3 (forced choice) Results

2 pairs: not replicated in both BJ and GZ.

3 pairs: GZ and BJ participants clearly differ (in statistical sense):

Exp 3 (forced choice) Results

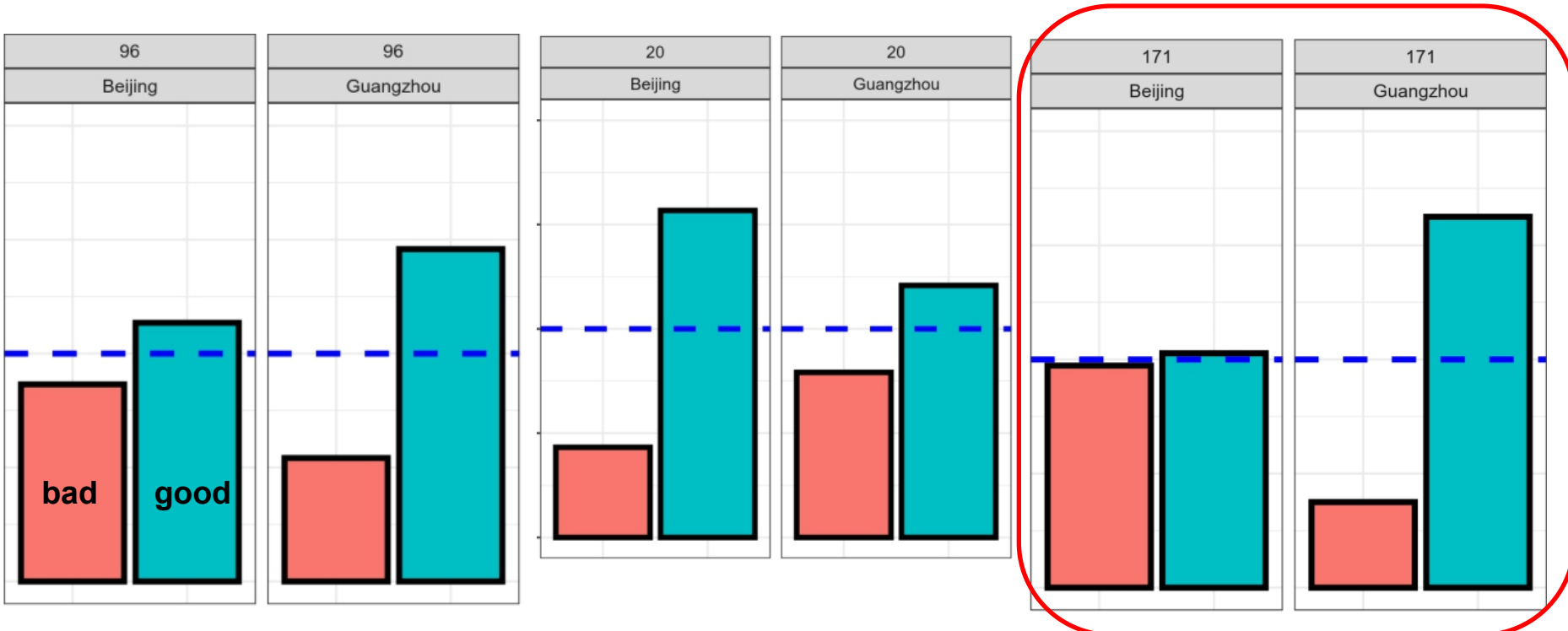
2 pairs: not replicated in both BJ and GZ

3 pairs: GZ and BJ participants clearly differ (in statistical sense):

Pair 96: bad: 他写过本书很有意思。 g: 他写过一本书很有意思 。

Pair 20: bad: 那个谣言是到处流传的。 g: 那个谣言是他已经病死了。

Pair 171: bad: 李奇笑下午，不是笑上午。 g: 李奇开下午，不是开上午。

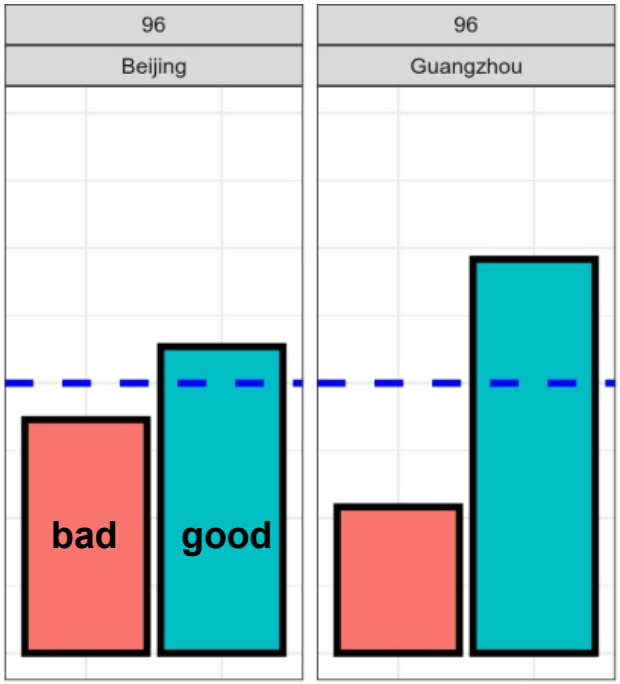


Exp 3 (forced choice) Results

Pair 96:

bad: 他写过本书很有意思。 he wrote CLS book very interesting

good: 他写过一本书很有意思。he wrote one CLS book very interesting

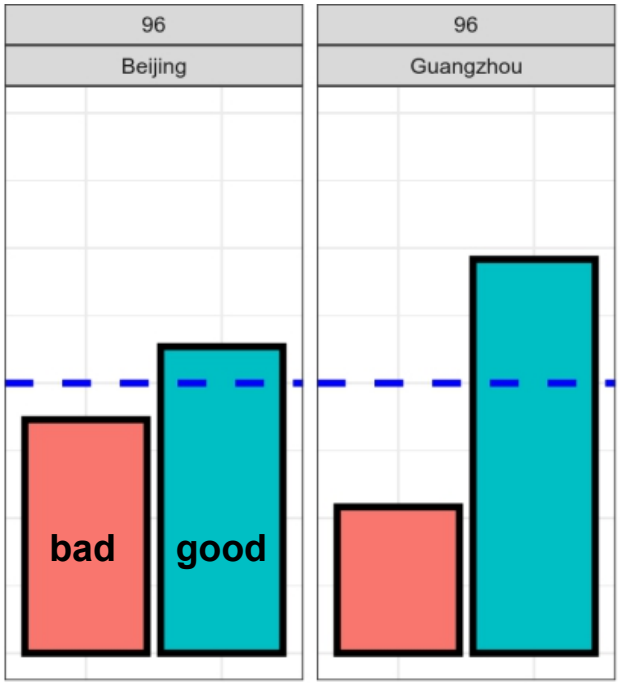


Exp 3 (forced choice) Results

Pair 96:

bad: 他写过本书很有意思。 he wrote CLS book very interesting

good: 他写过一本书很有意思。he wrote one CLS book very interesting



BJ participants more tolerant of omitting ‘one’

GZ participants like ‘one + classifier’ more

However, in Cantonese, ‘null + classifier’ is preferable.

→ Bilinguals very sensitive to L1/L2 boundary

Discussion

- Convergence rate:
  - Likert scale: 86%
  - Forced choice: (337-19-5)/337 = 93%
- Lower than Chinese textbook:89%,96% (Chen et al 2020)
  - Sentences in research articles are more controversial
- Lower than English: 95% (Sprouse et al 2013)
  - Discourse related pairs
  - A wider range of journals/papers
- What is grammar?
  - “pure” syntax vs. discourse
  - typologically different languages

Discussion

- Dialectal/language influence: Exists, but not too large
  - Beijing vs Guangzhou
  - Exp1: 26 out of 337 pairs
  - Exp3: only 3 pairs show sig. difference between two groups
- High overlap in judgments → they have same grammar for Mandarin
- GZ participants have clear boundaries between L1 and L2

Conclusion and future work

- Convergence rate comparable to, but lower than previous research on English, or Chinese textbook
- Dialectal difference exists, but not too large
- Author background does not play a role
- Chinese has no grammar?
  - It does!
  - But there may be more borderline cases
- Future work:
  - BJ Participants: Beijing Mandarin is different from Standard Mandarin
  - Testing specific syntactic phenomena

Acknowledgement

We thank Licen Liu, Yushu Wang, Xiaojie Gong, Carol Zheng, Qi Zhang, Xiaojing Zhao, Zihan Zhao, and many others for their help in data collection.

We also thank Zhong Chen and Yuhang Xu for help with the R script and Qualtrics setup.

This project is dedicated to Jiahui Huang.

Thank You!
Questions and comments are welcome!