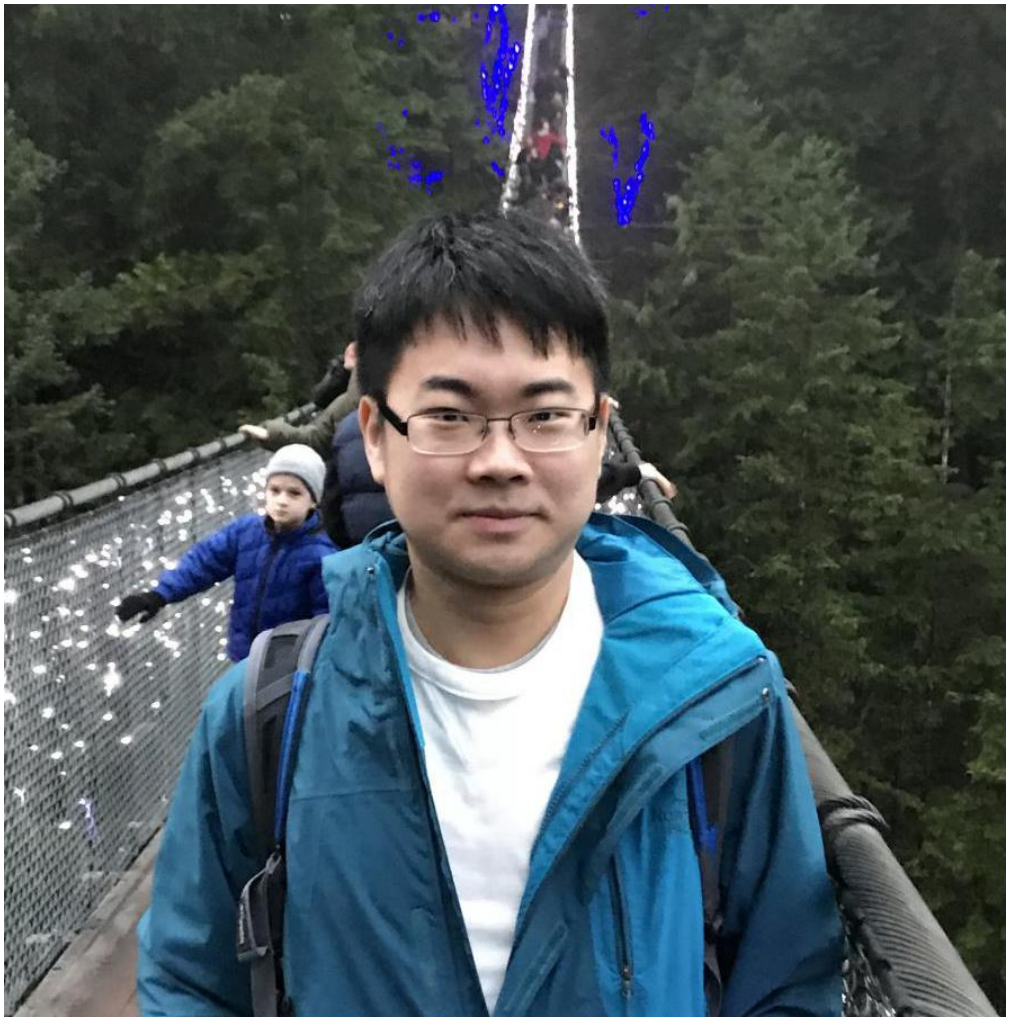
Replicating Acceptability Judgments from Syntax Papers in Chinese

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

1: Indiana University Bloomington
2: University of Pennsylvania
3: University of Washington
4: Brigham Young University

2021 June, NACCL 33

Dedicated to our dearest friend: Jiahui Huang



Background

- Much work has been done to examine whether the informal grammatical judgments in syntactic research can be replicated under experimental settings:

  - Sprouse et al (2012, 2013):
    - Acceptability judgments in **English** in syntax **textbook** (Adger 2003) and **journal** (LI) are valid and robust under more formal experimental settings
  - Chen et al (2020)
    - Acceptability judgments in **Chinese** in syntax **textbook** (Huang et al 2009) are valid and robust under experimental setting: convergence rate = 89.2%

The current study

- This study conducts a large-scale acceptability judgment experiment using sentences randomly sampled from academic **journal articles** on Chinese syntax.

- Research question

  Whether the informal grammatical judgments from journal papers on Chinese syntax can be replicated under a formal experimental setting?

(we use “grammaticality” and “acceptability” interchangeably)

Method

- **Obtaining our stimuli: data sampling**

Goals: wide coverage; primarily on a topic in syntax; minimal pairs

1. Select journals
    a. Peer-reviewed, high-quality, publish papers on Chinese syntax
    b. → 10 journals (wide coverage)
2. Find papers on Chinese syntax 2010-2020
    a. Standard Mandarin Chinese; not dialect; not archaic Chinese
    b. → 128 papers → sample → 68 papers
3. Find ungrammatical sentences
    a. Copy all examples (incl. footnote) to excel sheet → 7261 examples
    b. Sample 6 ungrammatical examples per paper → 397 sentences
4. Find minimal pairs
    a. Find/construct minimal pair
    b. Remove examples involving: anaphora, interpretation, prosody
    c. → 337 pairs

Method

- **Stimuli for judgment**

337 minimal pairs, 92 w/ constructed control sentence

journal	language	n papers	n sents	n pairs
Concentric: Studies in Linguistics	English	5	38	19
J of Chinese Ling	English	6	70	35
J of East Esian Ling	English	6	52	26
Linguistic Inquiry	English	5	48	24
Language & Linguistics	English	6	62	31
Lingua Sinica	English	6	66	33
Natural Lang and Ling Theory	English	6	58	29
Taiwan Journal of Linguistics	English	7	72	36
语言科学	Chinese	10	102	51
中国语文	Chinese	11	106	53
sum		**68**	**674**	**337**

Authors from mainland China, Taiwan, Hong Kong and elsewhere

In sum, representative samples

Method

- **Judgments collection**

- 674 sentences → randomly split into 6 lists
- No pair had both sentences in the same list, i.e., the grammatical and ungrammatical sentences for one pair will be in two different lists
- Online questionnaire distributed using Qualtrics
- Two catch trials interspersed in each list (Chen et al 2020)

    e.g., 这道题请直接选择3这个选项

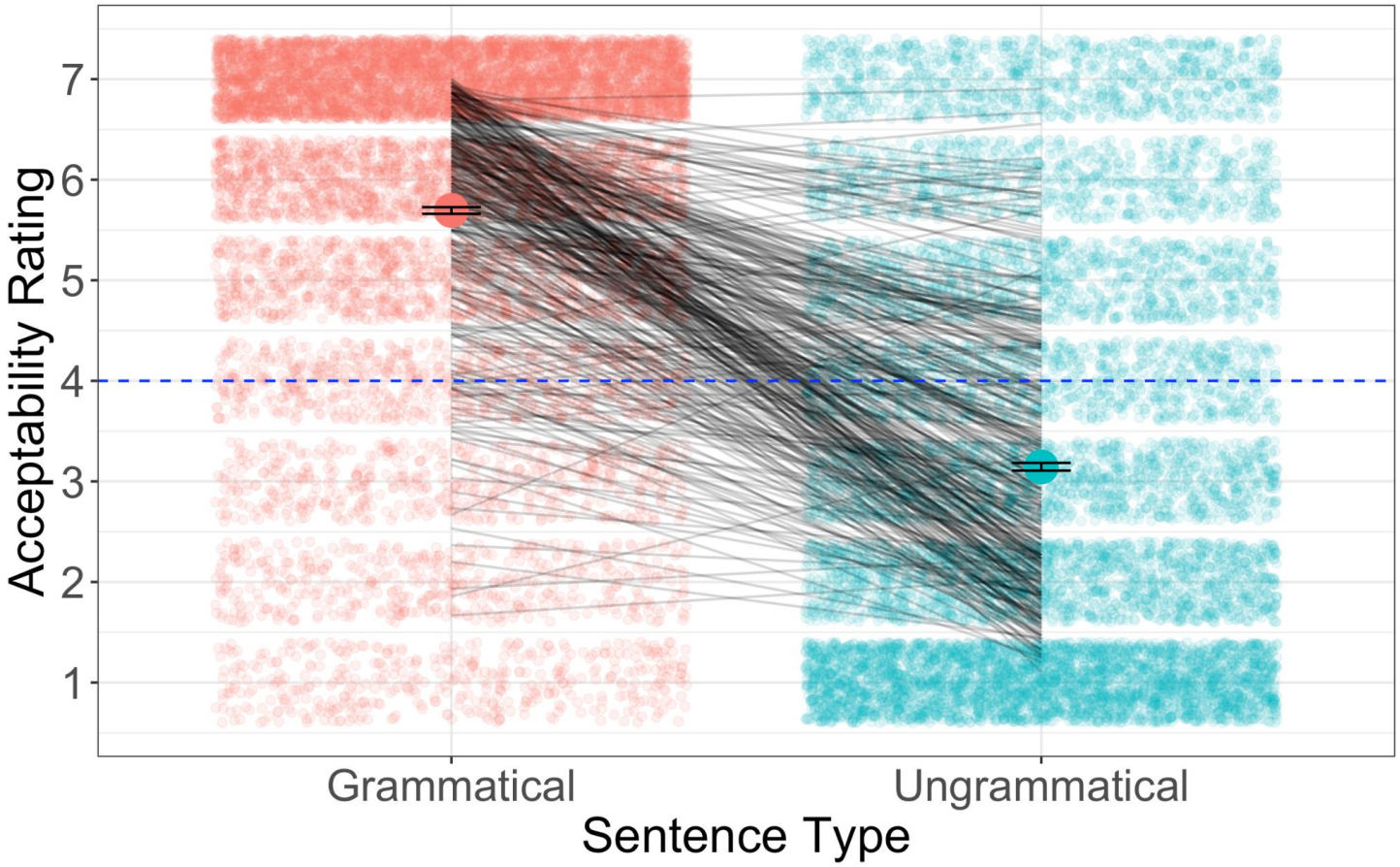
Method

- **Participants**

- 223 native speakers of Mandarin Chinese (Beijing Mandarin) were recruited to rate the naturalness of the sentences on a 7-point Likert scale
- Each participant was randomly assigned to one list
- 36 were excluded due to
  - completion time less than 5 minutes
  - failure to correctly answer the catch trials
  - spent more than 2 years outside BJ before age 18
- 187 participants included for statistical analysis
  - 142 female, 45 male, mean age=22, sd = 5.37

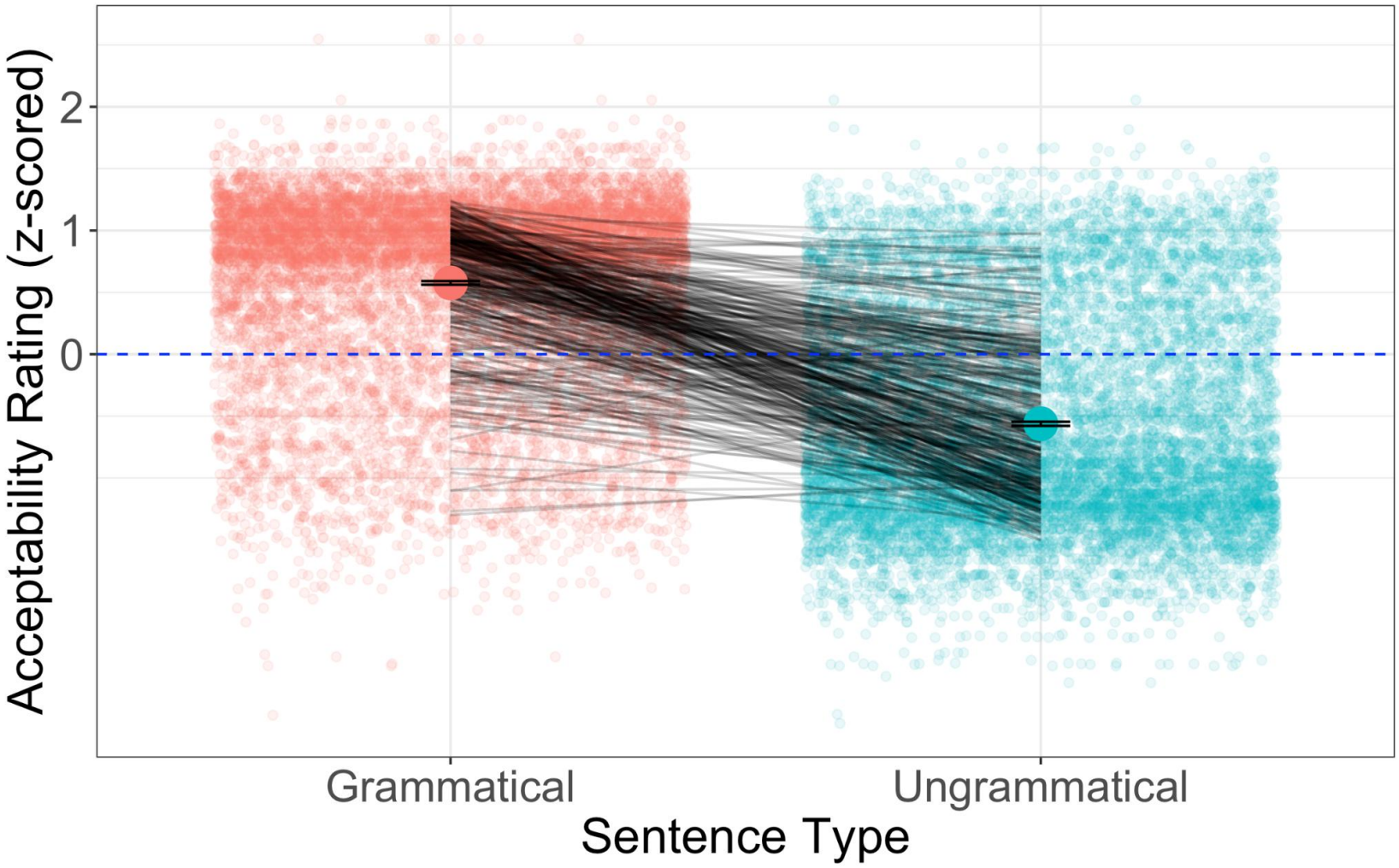
Results: mean rating

- Mean acceptability rating (raw scores)
- Grammatical sentences: 5.69
- Ungrammatical sentences: 3.14



Results: z-score transformed

- Mean acceptability rating (z-score)



Results: regression model

- A linear mixed-effects regression model was fit to the data

- Ratings_z ~ gram + (gram|participant) + (gram|item)

- Dependent variable: ratings (z-score transformed)
- Independent variable: original paper grammaticality judgement
- Random slopes: grammaticality by participant, grammaticality by item
- Ungrammatical items were rated significantly lower (β = -1.14, $P < 0.001$)

Results: t-tests

- Two-tailed paired t-tests for each pair
- **289 / 337 pairs replicated**
  - i.e., grammatical sentences rated significantly higher than ungram.
- 48 / 337 not replicated, 3 categories:
  - Grammatical < ungrammatical (sig): 4 pairs
  - Grammatical < ungrammatical (nonsig): 16 pairs
  - Grammatical > ungrammatical (nonsig): 28 pairs
- Convergence rate = 289/337 = 85.8%
  - c.f., 89.2% in Chen et al (2020)

Results

- Which pairs are not replicated?
- Go over all 337 pairs and found 32 pairs that are open to different interpretations or problematic (might be excluded)
  - *他居然给我们喝了三瓶酒 vs. 他居然给我喝了三瓶酒 (ambiguity)
  - *他认真的回 vs. 他认真的爬 (wrong spelling)
- Within 48 non-replicated pairs, 8 of them are in this category
- New convergence rate = (305-40)/305 = 86.9%
- Now, let's look at those non-replicated pairs!

Results

1. Grammatical < ungrammatical: sig (4 pairs), p-threshold = 0.05

Pair_id	Sentence	Journal	Gram	Mean Rating
12	谁，你认识，而谁，你又不认识？	CSL	g	4.4
12	什么东西你买了，什么东西还没买？	CSL	u	5.7
125	你吃这双筷子吧！	LL	g	1.9
125	你吃这把叉子吧！	LL	u	3.3
254	中国古代从来防止人口流动。	yykx	g	2.7
254	中国古代从来没有防止人口流动。	yykx	u	5.1
272	一个日本军官比比划划地讲着日本话。	yykx	g	5.3
272	一个日本军官比比划划地在讲什么？	yykx	u	5.9

Reminder: each sentence rated by at least 30 participants.

Results

2. Grammatical < ungrammatical: not sig (16 pairs)

Pair_id	Sentence	Journal	Gram	Mean Rating
19	他知道了你一直不肯告诉我的，我昨天自己听到的那个消息。	JCL	g	4.1
19	他知道了邓小平逝世，探险家抵达南极的消息。	JCL	u	4.8
25	我被批评了。	JCL	g	6.8
25	我被表扬了。	JCL	u	6.9
186	那只狐狸已经跑得我筋疲力尽了，可我还是追不上它。	NLLT	g	3.8
186	那只狐狸已经把我跑得筋疲力尽了，可我还是追不上它。	NLLT	u	4.8

Results

3. Grammatical > ungrammatical: not sig (28 pairs)

Pair_id	Sentences	Journal	Gram	Mean Rating
20	那个谣言是他已经病死了。	JCL	g	5.5
20	那个谣言是到处流传的。	JCL	u	4.7
23	这是你私自对那件事的批评。	JCL	g	3.7
23	这是你对那件事的私自批评。	JCL	u	3.1
60	张三看到某人，但我不知道是谁。	JEAL	g	4.3
60	张三看到某人，但我不知道谁。	JEAL	u	4.1
96	他写过一本书很有意思。	LI	g	6.0
96	他写过本书很有意思。	LI	u	5.0

Conclusion and future work

- In general, judgments in academic papers on Chinese syntax can be replicated under experimental setting.
- Age and BJ-Mandarin might play a role.

For the future

- A forced choice task to check whether the non-replicated pairs can be replicated.
- Looking at the effects of other factors.

Acknowledgement

We thank Licen Liu, Yushu Wang, Xiaojie Gong for their help in data collection.

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

Thank You!
Questions and comments are welcome!