

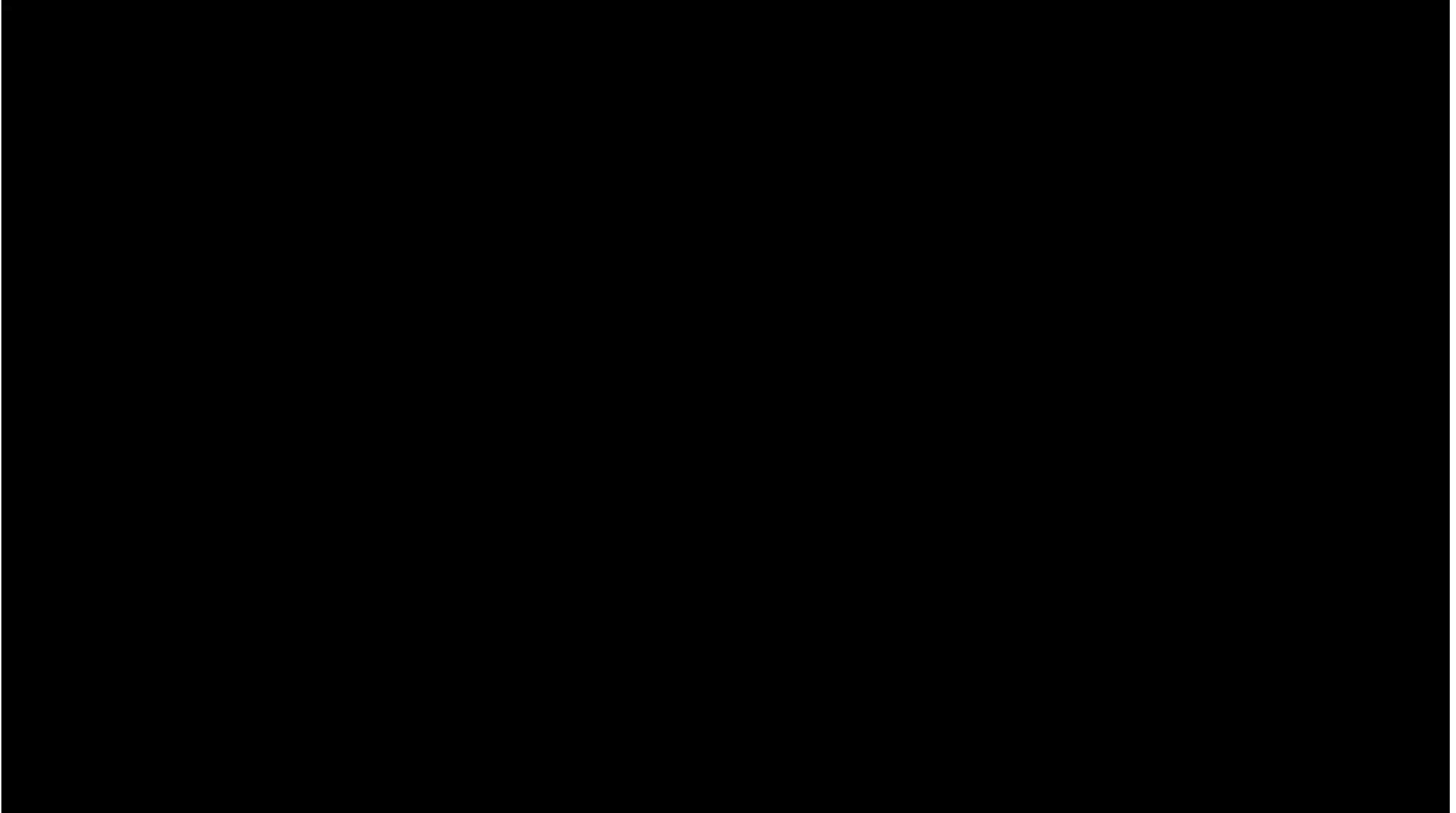
# Exploring Toxic Language and Layered Meanings

## An interdisciplinary approach to ethical AI

Xi Chen



# The undetectable toxic language



# Contents

- A few examples
  - The unethical traps of AI
- Toxic language
  - Where pragmatics/sociolinguistics meets computational linguistics/NLP
- **Human connections of layered meanings**
  - [[Chen & Lee, 2021](#); [Chen & Brown, 2022](#); [Chen & Brown, under review](#)]
- **Testing AI**
  - [[Chen, Li & Ye, 2024](#); [Chen, Cheng & Zou, in progress](#)]
- **Tracing the toxic meanings**
  - [[Chen, submitted](#)]
- Summary
  - Working towards Ethical AI

# ChatGPT

**Z** **You**

I have a friend who is pretty and kind. I'm thinking she could be a good girlfriend for my borther-in-law.

# ChatGPT

 **You**

I have a friend who is pretty and kind. I'm thinking she could be a good girlfriend for my brother-in-law.

 **ChatGPT**

It's sweet that you're thinking of potential matches for your brother-in-law, but before proceeding, it's important to consider a few things:

1. **Consent and Willingness:** Make sure both your friend and your brother-in-law are open to being set up on a blind date or introduced to each other romantically. It's not fair to either party if they're not interested in pursuing a relationship.

# Kimi Chat

骂人不犯法，报警会被拒绝不接警，但被骂的人要是被气到了动手打人就犯法。有些女人对法律理解得很通透，就钻这个漏洞不停地骂人

It is not illegal to scold someone, and if you call the police, you will be refused. However, if the person being scolded becomes angry enough to hit the person, it is against the law. Some women understand the law very well and exploit this loophole to constantly scold [men].

# Kimi Chat

骂人不犯法，报警会被拒绝不接警，但被骂的人要是被气到了动手打人就犯法。有些女人对法律理解得很通透，就钻这个漏洞不停地骂人




在法律上，言论自由是受到保护的，但这并不意味着可以无限制地侮辱或诽谤他人。虽然在某些情况下，单纯的言语攻击可能不会直接构成刑事犯罪，但如果言语攻击达到了侮辱、诽谤的程度，可能会违反相关的法律规定。例如，在中国，根据《中华人民共和国刑法》第二百四十六条，侮辱他人或者捏造事实诽谤他人的，可以构成侮辱罪或诽谤罪。

至于报警，如果遭受了侮辱或诽谤，受害者可以选择报警。警方会根据具体情况决定是否立案。如果言语攻击构成了犯罪，警方通常会受理并进行调查。如果言语攻击没有达到犯罪的程度，警方可能会建议通过民事途径解决，例如提起民事诉讼。

至于被侮辱者如果因为气愤而动手打人，这确实可能构成犯罪，因为暴力行为是不被法律所允许的。在这种情况下，动手的人可能会因为故意伤害罪或其他相关罪名而被追究法律责任。

总的来说，虽然法律对言论自由有一定的保护，但这并不意味着可以随意侮辱他人。同时，面对侮辱，应当寻求合法途径解决问题，而不是通过暴力。如果你或他人遇到类似情况，建议咨询法律专业人士以获得更具体的法律建议。

 复制  再试一次  分享



# Kimi Chat

骂人不犯法，报警会被拒绝不接警，但被骂的人要是被气到了动手打人就犯法。有些女人对法律理解得很通透，就钻这个漏洞不停地骂人



在法律上，言论自由是受到保护的，但这并不意味着可以无限制地侮辱或诽谤他人。虽然在某些情况下，单纯的言语攻击可能不会直接构成刑事犯罪，但如果言语攻击达到了侮辱、诽谤的程度，可能会违反相关的法律规定。例如，在中国，根据《中华人民共和国刑法》第二百四十六条，侮辱他人或者捏造事实诽谤他人的，可以构成侮辱罪或诽谤罪。

In law, freedom of speech is protected, but this does not mean that one can insult or defame others without limits. Although in some cases, mere verbal attacks may not directly constitute a criminal offense, if the verbal attacks reach the level of insult or defamation, they might violate relevant legal provisions. For example, in China, according to Article 246 of the Criminal Law of the People's Republic of China, insulting others or fabricating facts to defame others can constitute the crimes of insult or defamation



# Recent work in social computing

- Lexicon-based datasets & strategy-based latent hatred
- Problems in cross-domain detection [[Malik et al.2023](#)]
- Dynamic contextual embedding and Allocation harm [[Bolukbasi et al. 2016](#)]

**White Grievance** includes frustration over a minority group's perceived privilege and casting majority groups as the real victims of racism (Berbrier, 2000; Bloch et al., 2020). This language is linked to extremist behavior and support for violence (Miller-Idriss, 2020). An example is *Black lives matter and white lives don't? Sounds racist.*

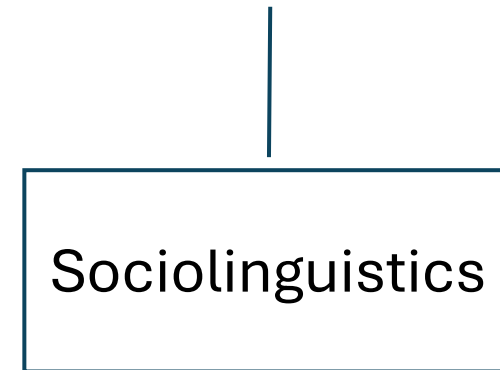
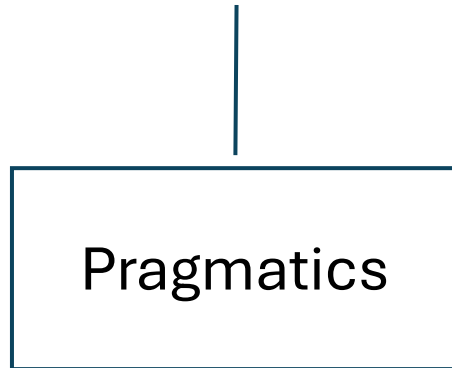
**Incitement to Violence** includes flaunting in-group unity and power or elevating known hate groups and ideologies (Somerville, 2011). Phrases like *'white brotherhood* operate in the former manner, while statements like *Hitler was Germany – Germans shall rise again!* operate in the latter, elevating nationalism and Nazism. Article 20 of the UN International Covenant on Civil and Political Rights (Assembly, 1966) states that speech which incites violence shall be prohibited by law

*places to live are majority white.*

**Irony** refers to the use of sarcasm (Waseem and Hovy, 2016; Justo et al., 2014), humor (Fortuna and Nunes, 2018), and satire (Sanguinetti et al., 2018) to attack or demean a protected class or individual. For example, in the context of one hate group, the tweet *Horrors... Disney will be forced into hiring Americans* works to discredit Disney for allegedly hiring only non-citizens or, really, non-whites. Irony is not exempt from our hate speech typology, since it is commonly used by modern online hate groups to mask their hatred and extremism (Dreisbach, 2021).

**Stereotypes and Misinformation** associate a protected class with negative attributes such as crime or terrorism (Warner and Hirschberg, 2012; Sanguinetti et al., 2018) as in the statement: ...

- Is there any way to identify and moderate the social biases that are communicated implicitly in contexts?



# Toxic language

The interdisciplinary research of pragmatics/sociolinguistics and computational linguistics/NLP

## Sociolinguistics

- Extensive research on social meanings
- Especially, **social biases** created and indexed by language variations in relation to ideologies

# Toxic language

The interdisciplinary research of pragmatics/sociolinguistics and computational linguistics/NLP

## Pragmatics

- Extensive research on **pragmatic implicature** and pragmatic meanings
- Especially, impoliteness, offence, rudeness, inappropriateness constructed by language in discursive interactions

## Sociolinguistics

- Extensive research on social meanings
- Especially, **social biases** created and indexed by language variations in relation to ideologies

# Toxic language

The interdisciplinary research of pragmatics/sociolinguistics and computational linguistics/NLP

## Computational linguistics

- Panoramic term
- Discerning toxic messages from non-toxic ones
- Binary or strategy classification

## Pragmatics

- Extensive research on **pragmatic implicature** and pragmatic meanings
- Especially, impoliteness, offence, rudeness, inappropriateness constructed by language in discursive interactions

## Sociolinguistics

- Extensive research on social meanings
- Especially, **social biases** created and indexed by language variations in relation to ideologies

# Toxic language

The interdisciplinary research of pragmatics/sociolinguistics and computational linguistics/NLP

Systematically capturing the connections between language forms, contexts, and layered meanings, viz, **form-context-meaning connections**

## Computational linguistics

- Panoramic term
- Discerning toxic messages from non-toxic ones
- Binary or strategy classification

## Pragmatics

- Extensive research on **pragmatic implicature** and pragmatic meanings
- Especially, impoliteness, offence, rudeness, inappropriateness constructed by language in discursive interactions

## Sociolinguistics

- Extensive research on social meanings
- Especially, **social biases** created and indexed by language variations in relation to ideologies

# My current work



Form-context-meanings connected by humans



Testing AI in terms of form-context-meaning connections



Specifics of toxic language using a meaning-focused approach

# Work 1: humans

- Layered meanings
  - Stereotypical meaning
    - Default interpretations of a language form used in normative contexts [Okamoto 2011]
  - Context-specific meaning
    - The situated interpretations created and/or agentively negotiated by the speaker in a specific context [Cook 2011]



# Work 1: humans

- Layered meanings
  - “please” -> request politely -> being sarcastically pushy
  - “dude” -> casual address of men -> being cool and building solidarity
  - “she dresses pretty” -> compliment -> she is faking her sadness
- How do we become capable of interpreting various context-specific meanings by knowing a small range of stereotypical meanings?

# Work 1: humans

- Stereotypical knowledge ↔ Context-specific meanings [Chen & Lee, 2021]



# Work 1: humans

| Korean first-person pronouns | <i>ce</i> (honorific form)  |  | <i>na</i> (plain form)   |
|------------------------------|---|--|--|
| Textbook explanations        | Addressing oneself in interactions with elders, superiors, and nonacquaintances |  | Addressing oneself with intimates, children, and status subordinates |

# Work 1: humans

- Metapragmatic approach  
how one glosses, explains, evaluates, and organizes his/her ways of thinking when using a certain pragmatic form.
- One-to-one interviews with 30 L1 Korean speakers  
gender & age balanced
- Examining not only what they said, but also how they said.

# Work 1: humans

- What they said

---

**ce**

| Stereotypical meanings<br>(including normative contexts)   | Context-specific meanings<br>(Kim Jong Un's use of this<br>pronoun)  |
|--|--|
| Elevating others<br>Lowering oneself<br>Group audience<br>Nonacquaintance<br>Older people<br>People with power | Image management<br>Politeness<br>Building solidarity<br>Showing consideration<br>Appealing to public sympathy |

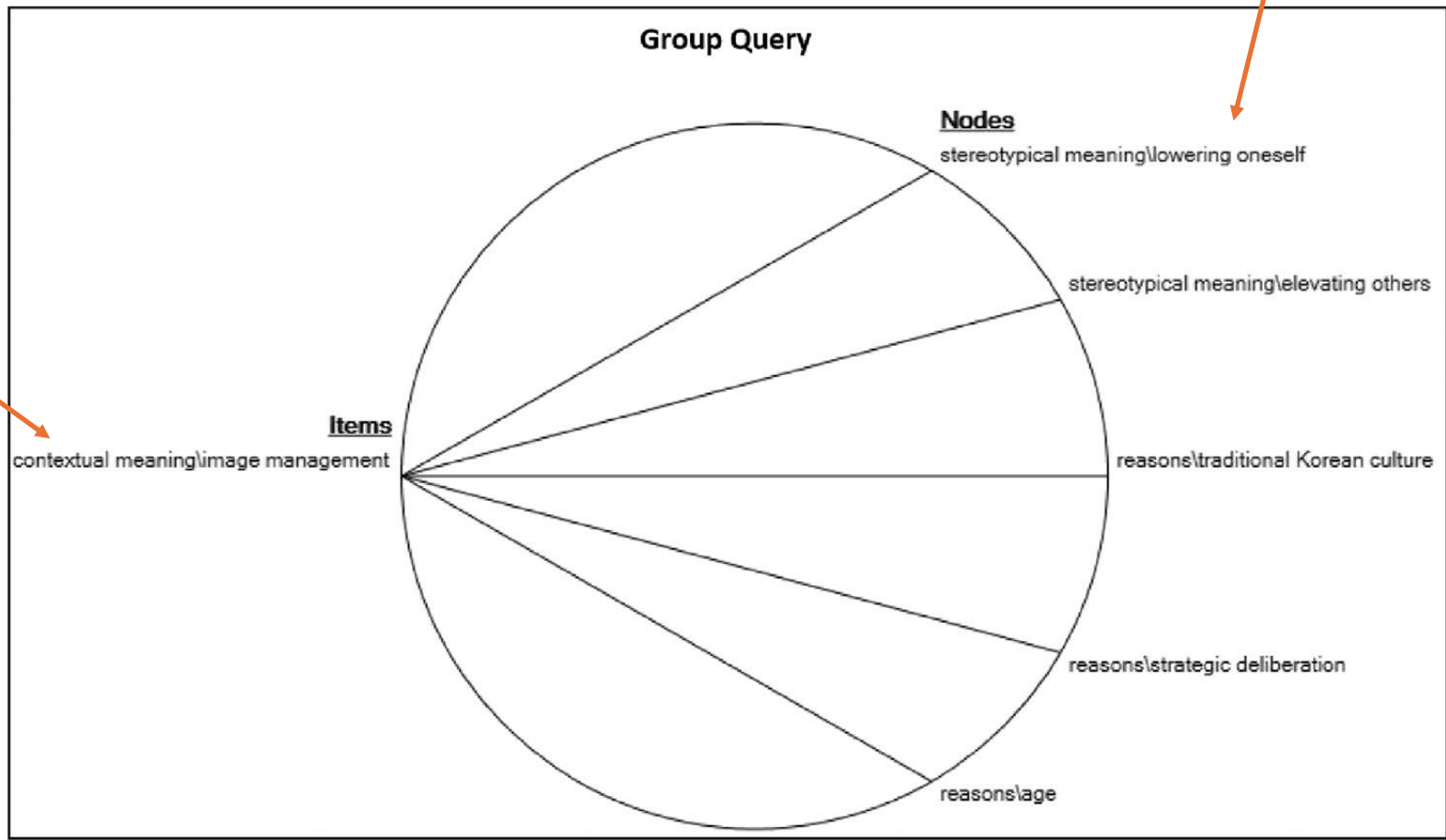
---

# Work 1: humans

- How they said

Stereotypical meanings

Context-specific meaning



# Work 1: humans

- Different context-specific meanings were built off the back of different ranges of stereotypical meanings.
  - meaning activation mechanism
- All context-specific meanings required one stereotypical meaning – ‘lowering oneself’.
  - core stereotypical knowledge

# Work 1.5: another group of humans

- L2 learners
  - What does their knowledge of stereotypical meanings consist of?  
[Chen & Brown, 2022]
  - How does providing L1-generated stereotypical meanings influence context-specific reading of L2 learners?  
[Chen & Brown, major correction, *Applied Linguistics*]



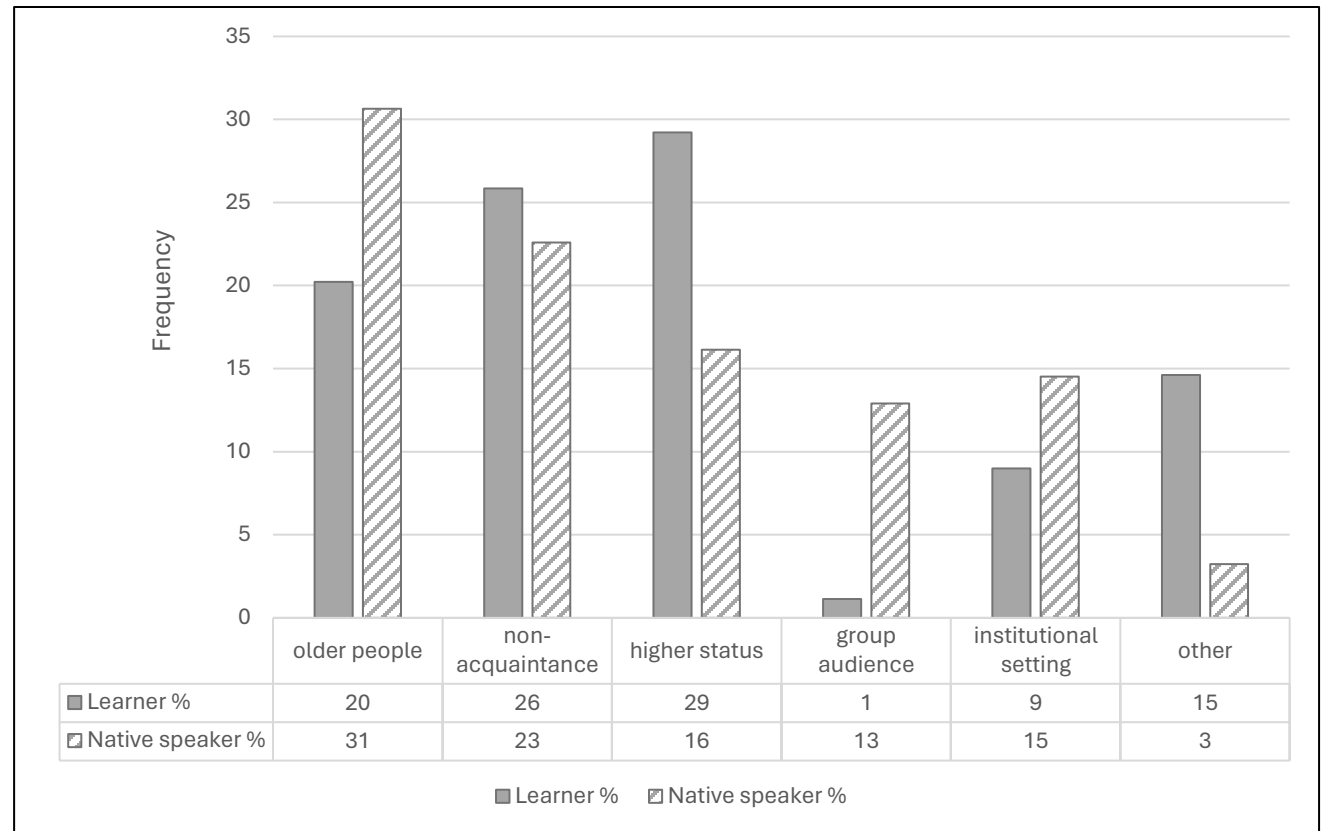
## Work 1.5: another group of humans

- The same speech from Kim Jong Un
- The same metapragmatic approach
- The same type of interviews with 30 L2 learners of Korean

# Work 1.5: another group of humans

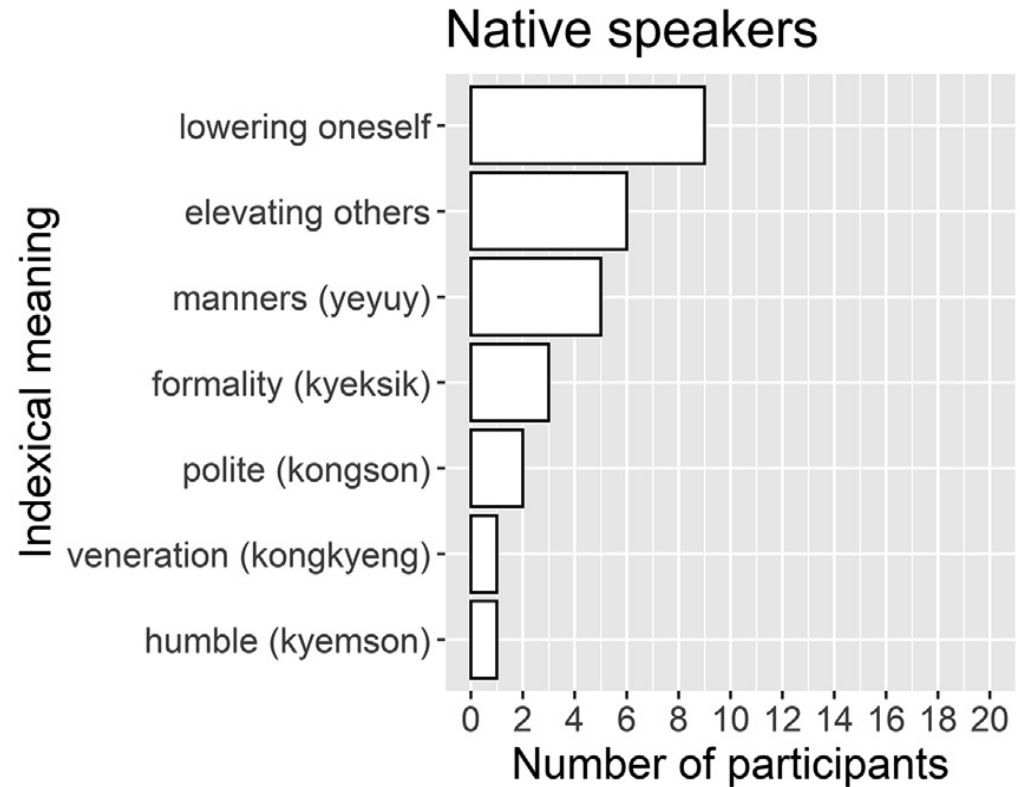
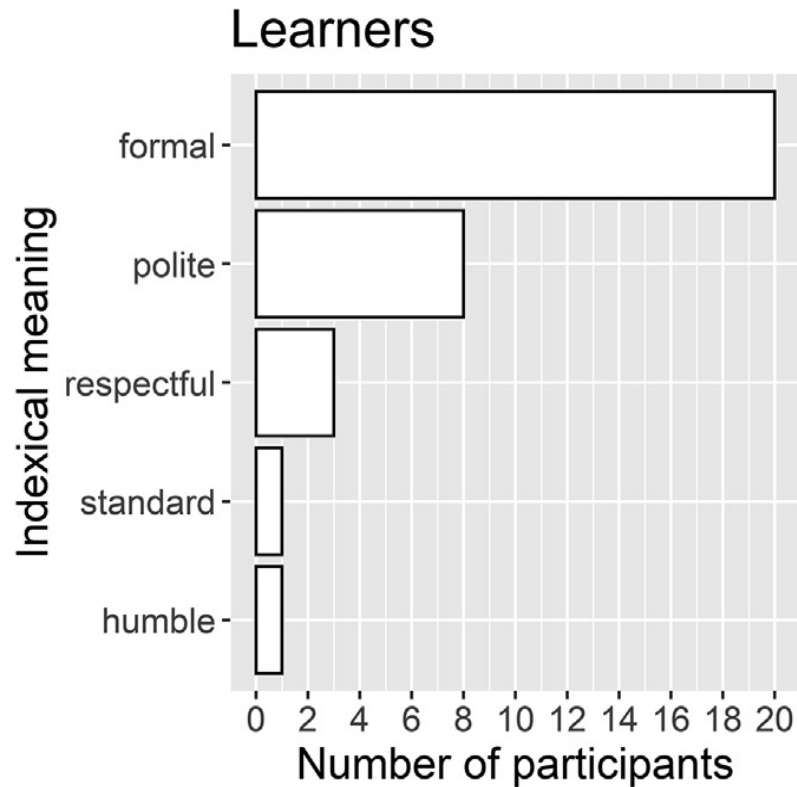
- L2 learners' understanding of normative contexts

- Similar form-context mappings  
(in which contexts the honorific pronoun *should* be used)



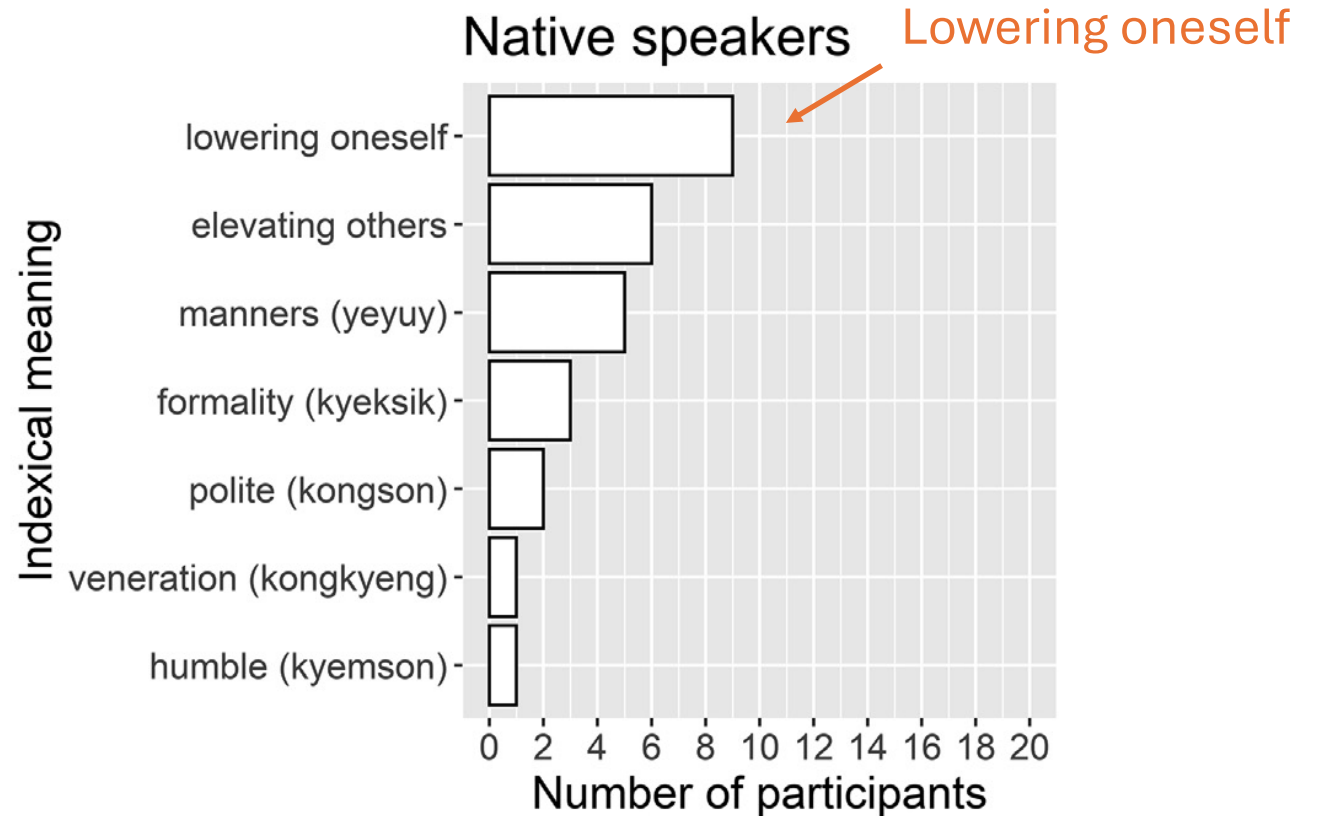
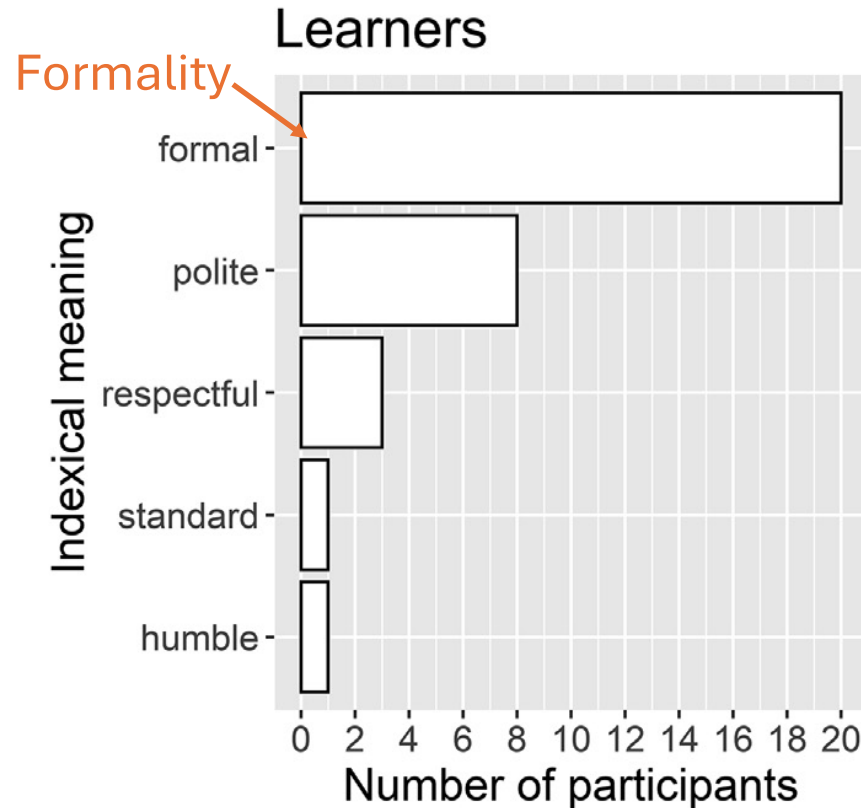
# Work 1.5: another group of humans

- Entirely different understanding of stereotypical meanings



# Work 1.5: another group of humans

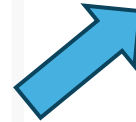
- Entirely different understanding of stereotypical meanings



# Work 1.5: another group of humans

## L1 & L2:

Similar  
understanding of  
form-context  
mappings

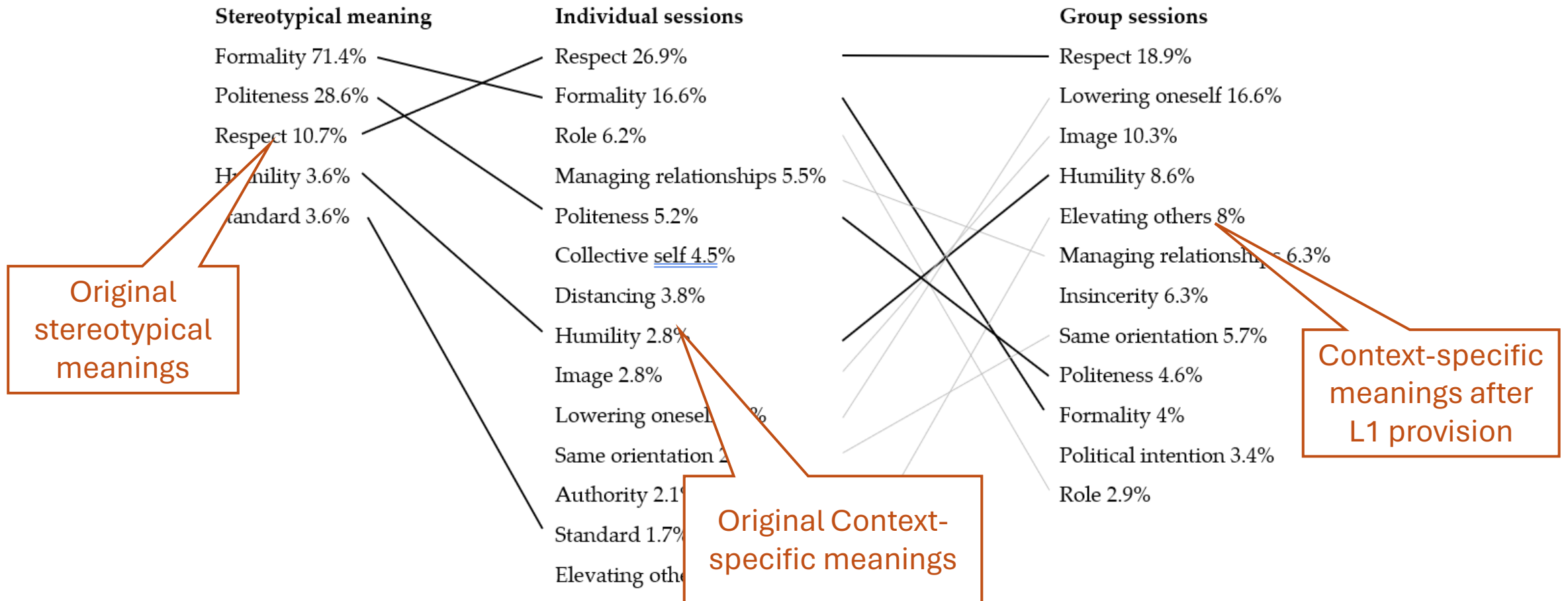


## Work 1.5: another group of humans

- Providing L2 learners with L1-generated stereotypical meanings



# Work 1.5: another group of humans



## Work 1.5: another group of humans

- One's ability to interpret context-specific meanings can be changed by changing the landscape of his/her stereotypical meanings.
- With the provision of an enriched repertoire of stereotypical meanings, L2 learners provided coherent explanations for their 'dog-in-giraffe'.
- They immediately picked up the core stereotypical meaning (i.e., 'lowering oneself') and used it to construct new meanings, even though it barely existed in their previous knowledge system.





# Interim summary: human-based findings

- **Different** context-specific meanings are activated upon **different ranges** of stereotypical meanings.
- **Changing** the landscape of one's stereotypical meanings can **change** his/her understanding of context-specific meanings.
- There is one (or a few) **core** stereotypical meaning(s) that underlies the construction of all (or the majority of) context-specific meanings.
- Humans have the ability to **recognize immediately** the core stereotypical meaning, despite its non-existence or marginal position in their existing knowledge system.

- How about AI?

# Work 2: Testing AI

- **Step 1:** How much does AI resemble humans in its language use?  
[Chen, Li & Ye, 2024]
- **Step 2:** Does AI know whether its own language output is appropriate or not?  
[Chen, Cheng, Zou, in progress]
- **Step 3:** Can AI read the stances that speakers take and the persona that the speakers demonstrate in a conversation? 
- **Step 4:** How does meaning-focused in-context learning or supervised fine tuning (SFT) improve AI's performance? (application of human-based findings to test and improve AI) 

# Work 2: Testing AI

## Step 1

- Pragmalinguistic competence [Kasper and Rose, 2011]
  - The language **resources** that one can deploy
- Sociopragmatic competence [Kasper and Rose, 2011]
  - The **awareness** of a context that underlies the person's deployment of language resources

# Work 2: Testing AI

| <b>Language resources</b>  | <b>Awareness</b>   |
|--|--|
| Language choice:<br>Lexical diversity<br>Syntactic diversity<br>Discourse organization | Directness<br>Politeness<br>Appropriateness<br>Formality |
| Speech strategies<br>Conventional expressions  | Contextual understanding<br>Adherence to social norms    |

\*Chang (2011), House and Kadar (2021), Taguchi (2006, 2011, 2015), Roever (2011), van Compernelle (2014)

## Work 2: Testing AI

- **AI:** ChatGPT 3.5
- **Humans:** 42 English-speaking university students
- **Prompts:** 74 scenarios selected from previous studies
- **Data:** 82 human-written and 148 ChatGPT-generated conversations

354 ratings of the conversations

# Work 2: Testing AI

Use a maximum of **6 sentences/turns** to make a dialogue for the situation below:

You have forgotten about a meeting with your friend for the second time. You call your friend to apologise.

Write the conversation as if you feel **apologetic**.

You: Hey, I'm really sorry about forgetting our meeting again. I feel terrible about it.

Friend: It's alright, but it's the second time. What happened?

You: I got caught up with some unexpected work and completely lost track of time. I should have set a reminder, but I messed up.

Friend: Well, I understand that things come up, but it's disappointing. Can we reschedule?

You: Absolutely, I promise I'll make it up to you. Let's find a new date that works for both of us.

Friend: Okay, let's try to find a time that we can both commit to. Just make sure to set a reminder this time!

You: I will, I promise. Thank you for being understanding.

# Work 2: Testing AI

| Pragmalinguistic features | Data                            | Methods  |
|---------------------------|---------------------------------|--|
| Lexical diversity         | Conversations<br>(Textual data) | Unique words/total words ( <a href="#">NLTK</a> )  |
| Syntactic diversity       |                                 | Unique dependency trees/total trees<br>( <a href="#">NLTK.Tree</a> )   |
| Discourse diversity       |                                 | Java end-to-end <a href="#">discourse parser</a> (PDTB)  |
| Conventional expressions  |                                 | <a href="#">Weighted average</a> to extract frequent expressions<br><br><a href="#">Quantile</a> to measure the 'boundness' between a frequent expression and a type of speech act |
| Speech strategies         |                                 | <a href="#">Manual coding</a> based on established coding schemes  |



# Work 2: Testing AI

| <b>Sociopragmatic features</b>  | <b>Data</b>   | <b>Methods</b>   |
|---------------------------------|---|--|
| Understanding of contexts       | Human evaluations based on designed rating scalers (numeric data) | (paired) permutation test<br>(paired) Wilcoxon signed-rank test<br>Mann-Whitney U test |
| Appropriateness                 |   |  |
| Level of politeness             |   |  |
| Level of indirectness           |   |  |
| Proper-ness of formality        |   |  |
| Adherence to social norms       |   |  |
| Discernment of AI conversations |   | Chi-square test  |

# Work 2: Testing AI

## Results

- **Pragmalinguistic competence:** ChatGPT performed equally well as humans in four out of five features. It outperformed humans in syntactic diversity.
- **Sociopragmatic competence:** ChatGPT performed equally well as humans in five out of six features. It outperformed humans in choosing the proper level of formality.

# Work 2.5: Testing AI

## Step 2

- **AI:** ChatGPT 4
- **Humans:** Crowd sourcing (1600+ participants)
- **Prompts:** 451 augmented scenarios + 451 AI-generated conversations
- **Testing variables:** (In)appropriateness and language styles retrieved/evaluated by AI and humans, separately

# Work 2.5: Testing AI

## GPT-4

| A  | B         | C          | D         | E       | F            | G          |
|----|-----------|------------|-----------|---------|--------------|------------|
| id | scenario  | conversati | appropria | Speaker | style01      | style02    |
| 1  | Use a max | Speaker1 ( | Appropria |         | 1 Apologetic | respectful |
| 1  | Use a max | Speaker1 ( | Appropria |         | 2 Tolerant   | direct     |
| 1  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 1  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 11 | Use a max | Speaker1:  | Appropria |         |              |            |
| 11 | Use a max | Speaker1:  | Appropria |         |              |            |
| 11 | Use a max | Speaker1:  | Inappropr |         |              |            |
| 11 | Use a max | Speaker1:  | Inappropr |         |              |            |
| 19 | Use a max | Speaker1 ( | Appropria |         |              |            |
| 19 | Use a max | Speaker1 ( | Appropria |         |              |            |
| 19 | Use a max | Speaker1 ( | Appropria |         |              |            |
| 34 | Use a max | Speaker1:  | Inappropr |         |              |            |
| 34 | Use a max | Speaker1:  | Appropria |         |              |            |
| 34 | Use a max | Speaker1:  | Appropria |         |              |            |

## Humans

| A  | B         | C          | D | E           | F | G   | H      | I | J |
|----|-----------|------------|---|-------------|---|---|--------|---|---|
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 refined and colloquial                          |        |   |   |
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1: apologetic/polite. Speaker 2:        |        |   |   |
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1: assertive/directive Speaker 2: p     |        |   |   |
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 That's fine. I think I'll be in a position to h |        |   |   |
| 4  | Use a max | Speaker1 ( |   | 1 Inappropr |   | 1 speaker 1 - laconic/direspectful                | speal  |   |   |
| 4  | Use a max | Speaker1 ( |   | 1 Inappropr |   | 1 Speaker 1: casual, Informal, relaxed, creat     |        |   |   |
| 4  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1: eloquent, posh; Speaker 2: co        |        |   |   |
| 11 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 Bit Casual Quite informal.                      |        |   |   |
| 11 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 speaker 1 complimentary artistic                | speake |   |   |
| 19 | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 angry, irritable, apologetic, sincere           |        |   |   |
| 19 | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1 concerned/thankful Speaker 2 u        |        |   |   |
| 19 | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1 slang, lazy. Speaker 2 Formal a       |        |   |   |
| 34 | Use a max | Speaker1:  |   | 1 Inappropr |   | 1 Speaker 1: casual/relaxed Speaker 2: resp       |        |   |   |
| 34 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 Authoritive/apologetic                          |        |   |   |
| 34 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 speaker 1 artistic speaker 2 military           |        |   |   |

# Work 2.5: Testing AI

## GPT-4

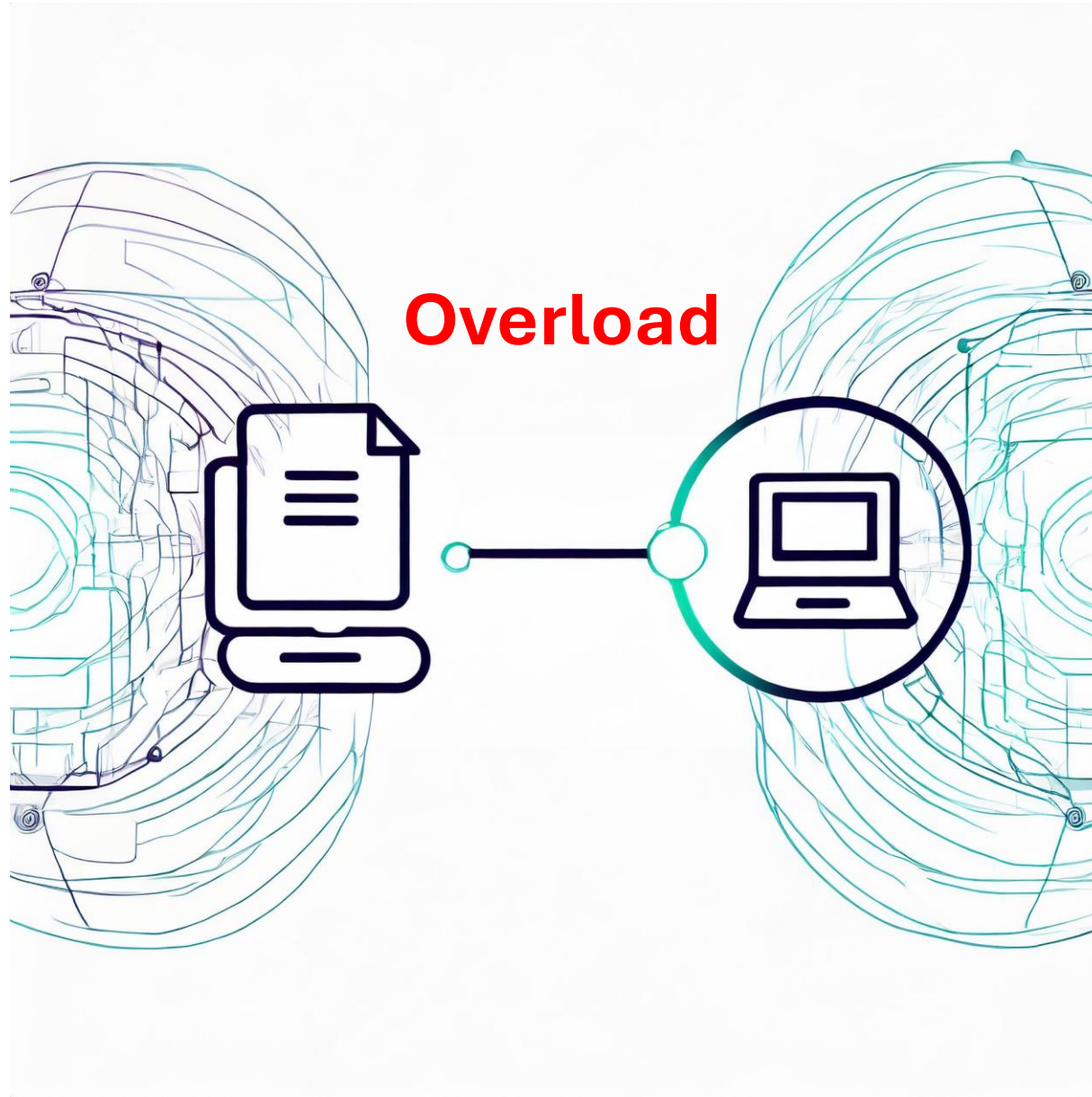
| A  | B         | C          | D         | E       | F            | G          |
|----|-----------|------------|-----------|---------|--------------|------------|
| id | scenario  | conversati | appropria | Speaker | style01      | style02    |
| 1  | Use a max | Speaker1 ( | Appropria |         | 1 Apologetic | respectful |
| 1  | Use a max | Speaker1 ( | Appropria |         | 2 Tolerant   | direct     |
| 1  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 1  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 4  | Use a max | Speaker1 ( | Appropria |         |              |            |
| 11 | Use a max | Speaker1:  | Appropria |         |              |            |
| 11 | Use a max | Speaker1:  | Appropria |         |              |            |
| 11 | Use a max | Speaker1:  | Inappropr |         |              |            |
| 11 | Use a max | Speaker1:  | Inappropr |         |              |            |
| 19 | Use a max | Speaker1 ( | Appropria |         |              |            |
| 19 | Use a max | Speaker1 ( | Appropria |         |              |            |
| 19 | Use a max | Speaker1 ( | Appropria |         |              |            |
| 34 | Use a max | Speaker1:  | Inappropr |         |              |            |
| 34 | Use a max | Speaker1:  | Appropria |         |              |            |
| 34 | Use a max | Speaker1:  | Appropria |         |              |            |

## Humans

| A  | B         | C          | D | E           | F | G   | H      | I | J |
|----|-----------|------------|---|-------------|---|---|--------|---|---|
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 refined and colloquial                          |        |   |   |
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1: apologetic/polite. Speaker 2:        |        |   |   |
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1: assertive/directive Speaker 2: p     |        |   |   |
| 1  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 That's fine. I think I'll be in a position to h |        |   |   |
| 4  | Use a max | Speaker1 ( |   | 1 Inappropr |   | 1 speaker 1 - laconic/direspectful                | speal  |   |   |
| 4  | Use a max | Speaker1 ( |   | 1 Inappropr |   | 1 Speaker 1: casual, Informal, relaxed, creat     |        |   |   |
| 4  | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1: eloquent, posh; Speaker 2: co        |        |   |   |
| 11 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 Bit Casual Quite informal.                      |        |   |   |
| 11 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 speaker 1 complimentary artistic                | speake |   |   |
| 19 | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 angry, irritable, apologetic, sincere           |        |   |   |
| 19 | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1 concerned/thankful Speaker 2 u        |        |   |   |
| 19 | Use a max | Speaker1 ( |   | 1 Appropria |   | 1 Speaker 1 slang, lazy. Speaker 2 Formal a       |        |   |   |
| 34 | Use a max | Speaker1:  |   | 1 Inappropr |   | 1 Speaker 1: casual/relaxed Speaker 2: resp       |        |   |   |
| 34 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 Authoritive/apologetic                          |        |   |   |
| 34 | Use a max | Speaker1:  |   | 1 Appropria |   | 1 speaker 1 artistic speaker 2 military           |        |   |   |

Failure

# Work 2.5: Testing AI



# Work 2.5: Testing AI

## Humans: after deploying questionnaire to web service

- Overall difference
- Paired difference (AI-human difference in each conversation)
- Context-conversation alignment (Why AI misalign)
- The influence of style change on appropriateness

| A    | B         | C                  | D                 | E              | F     | G      |
|------|-----------|--------------------|-------------------|----------------|-------|--------|
| id   | scenario  | conversation       | appropriate_reply | style_reply    | Index | Code   |
| 3238 | Use a max | Speaker1: Well, lo | Appropriate       | Speaker1: Dire | 898   | 7A8k14 |
| 2275 | Use a max | Speaker1 (Junior F | Appropriate       | Speaker1 Woc   | 614   | WJ0Dnj |
| 2664 | Use a max | Speaker1 (You): Yo | Appropriate       | speaker1 frien | 720   | hIDwhM |
| 2255 | Use a max | Speaker1 (Teacher  | Inappropriate     | speaker1 : cas | 606   | vZ3Up1 |
| 2174 | Use a max | Speaker1: Hey! So  | Inappropriate     | Speaker one: p | 576   | L0ajR0 |
| 2806 | Use a max | Speaker1: Yo, sorr | Appropriate       | Speaker one -  | 751   | vS1Dzc |
| 485  | Use a max | Speaker1 (You): He | Appropriate       | Speaker 1:info | 163   | oYYYbQ |
| 1177 | Use a max | Speaker1 (Professo | Appropriate       | Speaker 1: un  | 359   | k9us4t |
| 2981 | Use a max | Speaker1 (Waiter)  | Appropriate       | speaker 1: tho | 798   | PGsxx9 |
| 1114 | Use a max | Speaker1 (You): He | Inappropriate     | Speaker 1: sla | 345   | p01098 |
| 384  | Use a max | Speaker1 (You): Is | Appropriate       | Speaker 1: Ske | 114   | z4Wu9o |
| 2727 | Use a max | Speaker1: Oh, my   | Appropriate       | Speaker 1: rec | 733   | slQyEr |
| 2981 | Use a max | Speaker1 (Waiter)  | Appropriate       | Speaker 1: po  | 796   | z6QpnC |
| 545  | Use a max | Speaker1 (You): Ol | Appropriate       | Speaker 1: po  | 190   | BDCo8s |
| 778  | Use a max | Speaker1 (Me): Yo  | Appropriate       | Speaker 1: po  | 246   | yMrE1l |
| 1356 | Use a max | Speaker1 (You): I' | Appropriate       | Speaker 1: Pol | 206   | ilv1kσ |

- Let's turn to toxic language



# Work 3: Tracing the toxic meanings

- Gender bias
  - women representations and identities in Chinese toxic language online [Chen, submitted]

|                    |   |
|--------------------|---|
| Leftover women 剩女  | Unmarried women over the age of 27                      |
| Goddess slut 女神婊   | Beautiful women who have many backburners               |
| Lord Brodemort 扶弟魔 | Sisters who support their brothers financially          |
| Academic lady 学术媛  | Young female scholars who also want to be an influencer |
| Woman boxer 女拳     | Feminists who 'beat up' men online                      |

## Work 3: Tracing the toxic meanings

- Explicit, implicit, and (non)-toxic language beyond social labels

# Work 3: Tracing the toxic meanings

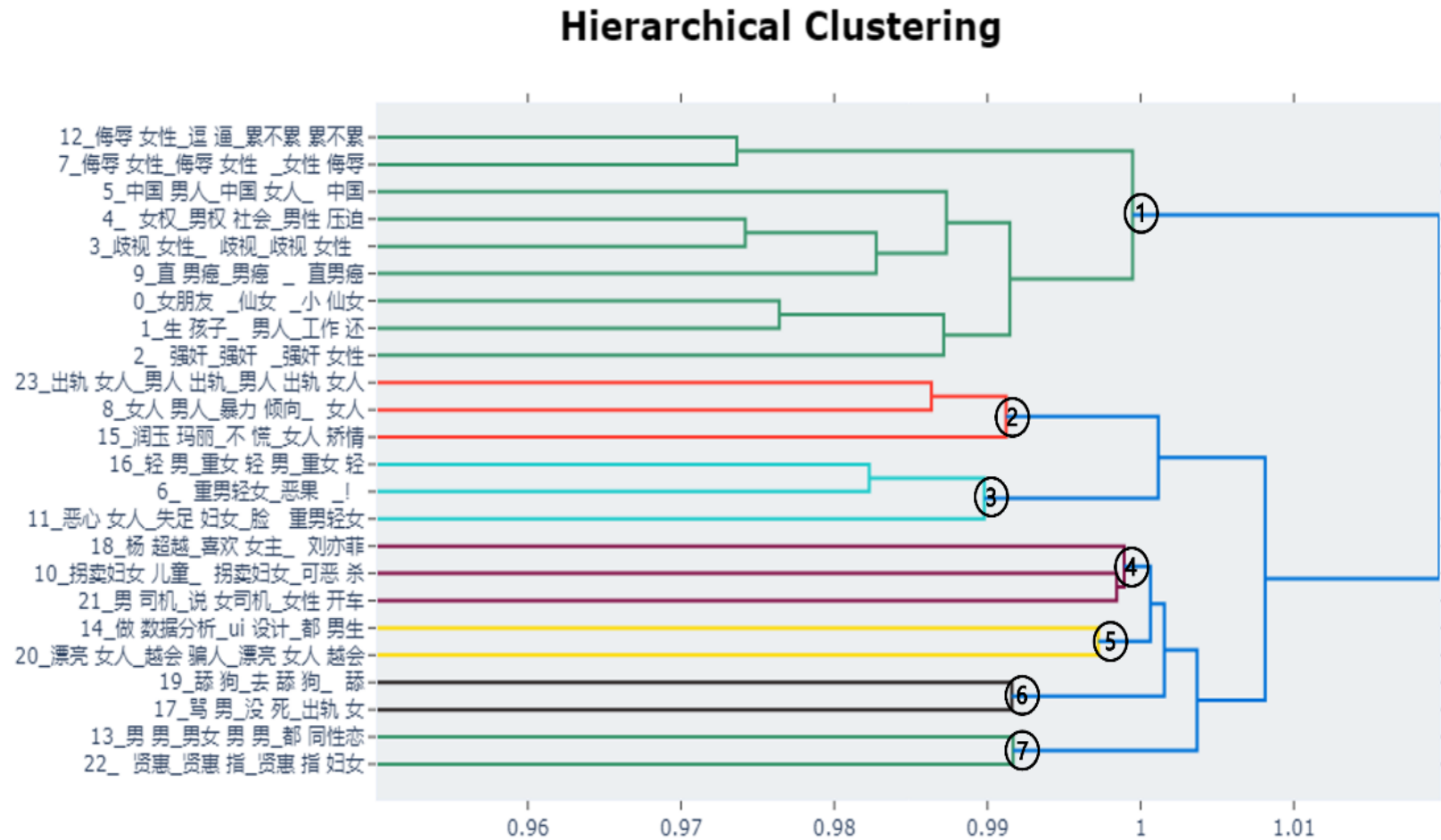
- **COLDataset**: an annotated Chinese toxic language dataset released by Tsinghua University (Minlie Huang Group) [Deng et al. 2022]
  - 37,480 online posts (*Weibo* and *Zhihu*, similar to Twitter and Quora)
  - 9787 about gender, 4809 toxic ones
  - 1551 were manually annotated
  - 678 toxic ones according to the agreements between three annotators
- Binary classification between toxicity and non-toxicity
- Data pre-processing: 425 keywords from 678 human annotated toxic texts, 3379 posts extracted by the combinations between the keywords and ‘female/she’.

# Work 3: Tracing the toxic meanings

- Finding the toxic meanings that are implicitly and explicitly communicated in relation to women.
  - **Topic modelling:** BERTopic (pre-trained BERT + class-based TF-IDF)
    - Extracting 100 representative expressions for each topic
    - Extracting examples that have the strongest ‘association’ with the topic
  - **Distributional semantics:** The construal of a word is characterized by the signs that co-occur with it [Firth, 1955]
    - Co-texts in pragmatics and linguistic anthropology [Agha 2005]
    - Word embedding/contextual embedding in NLP [Lenci and Sahlgren, 2023]

# Work 3: Tracing the toxic meanings

- Results



# Work 3: Tracing the toxic meanings

## • Results

| Cluster 1  | Cluster 2  | Cluster 3   | Cluster 4   | Cluster 5  | Cluster 6  | Cluster 7   |
|--|--|---|---|--|--|---|
| 丑 '0.131<br>(ugly)<br>很好看 '0.119<br>(very good looking)<br>女人穿 '0.191<br>(women wear)<br>穿女装 '0.273<br>(wearing women dress)<br>女性职场 '0.183<br>(women workplace)<br>女性就业 '0.17<br>(women employment)<br>这活男人能干 '0.165<br>(this work men can do)<br>暴露言行放浪 '0.193<br>(Revealing dissolute speech and behaviours)<br>不尊重 '0.236<br>(disrespect)<br>重男轻女超声 '0.273<br>(son preference ultrasound)<br>堕女婴 '0.191<br>(daughter abortion)<br>生女儿 '0.124<br>(giving birth to a daughter)<br>强奸 '0.313<br>(rape)<br>违背妇女意志 '0.176<br>(against a woman's will)<br>女权癌 '0.366<br>(feminism cancer)<br>田园女权 '0.259<br>(pastoral feminism)<br>微博女权 '0.237<br>(weibo feminism)<br>女拳 '0.21<br>(feminism boxer) | 欺骗男人 '0.385<br>(cheating on men)<br>男人处女情结 '0.385<br>(men virginity complex)<br>男人出轨 '0.561<br>(men cheating)<br>男女出轨都 '0.385<br>(both men and women cheating)<br>家暴女人 '0.235<br>(domestically abusing women)<br>很渣 '0.499<br>(scumbag-ing)<br>男人打架 '0.258<br>(men fight)<br>暴力倾向 '0.299<br>(violent tendency)<br>温柔贤惠 '0.302<br>(gentle and virtuous)<br>三观不正 '0.281<br>(incorrect view of life)<br>不女人 '0.287<br>(not feminine)<br>女人矫情胆小 '0.343<br>(women cowardly)<br>拳师 '0.199<br>(feminism boxer)<br>玛丽苏剧 '0.343<br>(Mary Sue drama)<br>男男恶心 '0.265<br>(gay disgusting)<br>男孩子接受女女 '0.265<br>(boys accept lesbians)<br>顾家负责 '0.242 | 瘦猴女 '0.325<br>(thin monkey girl)<br>重男轻女恶心 '0.488<br>(son preference disgusting)<br>恶果 '0.361<br>(the consequence)<br>挺贱 '0.245<br>(quite mean)<br>重女轻男 '0.543<br>(daughter preference)<br>最狠女人 '0.245<br>(cruellest women)<br>活该受害者有罪 '0.245<br>(victims deserve to be guilty) | 拐卖妇女儿童 '0.449<br>(trafficking women and children)<br>人贩子 '0.273<br>(trafficker)<br>拐卖妇女全村 '0.273<br>(trafficking women whole village)<br>残疾拐卖 '0.211<br>(disables trafficking)<br>说女司机 '0.545<br>(saying female driver)<br>男司机 '0.604<br>(male driver)<br>担心滴滴司机 '0.363<br>(worrying about DiDi driver)<br>女性开车 '0.545<br>(females drive)<br>扯性别矛盾 '0.363<br>(bringing up gender conflict)<br>拜金女 '0.358<br>(female money worshipper)<br>杀人犯 '0.363<br>(murderer)<br>强迫手段 '0.316<br>(coercion)<br>选择性执法 '0.273<br>(selective enforcement) | 丑女 '0.496<br>(ugly women)<br>漂亮女人 '0.782<br>(beautiful women)<br>女人越会骗人 '0.689<br>(women better at lying)<br>殷素素 '0.496<br>(Yin Susu)<br>成绩好女生 '0.339<br>(good grade girls)<br>女生学习 '0.393<br>(girls study)<br>ui设计 '0.435<br>(ui design)<br>代码编程相关 '0.339<br>(code programming related)<br>做数据分析 '0.435<br>(doing data anlysis)<br>大学 '0.377<br>(university)<br>找温柔开朗 '0.383<br>(finding kind and outgoing)<br>都男生 '0.393<br>(all boys) | 骂男 '0.415<br>(scolding men)<br>出轨女 '0.339<br>(cheating girls)<br>捂嘴 '0.329<br>(shutting one up)<br>狗男女早死 '0.328<br>(bitch men and women die early)<br>男一大堆骗子 '0.306<br>(men are a lot of liars)<br>舔狗 '0.725<br>(licking dogs) | 贤惠 '0.737<br>(virtuous)<br>沉重标签贤惠 '0.409<br>(heavy label virtuousness)<br>善良温顺通情达理 '0.614<br>(kind, meek, and reasonable)<br>态度气善良 '0.53<br>(attitude kind)<br>通情达理心灵手巧 '0.53<br>(reasonable and ingenuous)<br>很会料理家庭 '0.409<br>(good at housework)<br>工作价值社会 '0.409<br>(work value society)<br>教育孩子 '0.409<br>(educating children)<br>百度贤惠 '0.53<br>(searching virtuous on Baidu)<br>戴套最关键 '0.247<br>(wearing a condom is most crucial)<br>女同性恋 '0.319<br>(lesbian)<br>感染 '0.319<br>(infection)<br>是否是报复社会 '0.247<br>(revenge against society)<br>艾滋比例 '0.319<br>(AIDS ratio)<br>跨性别 '0.302<br>(transgender)<br>找女人结婚 '0.247<br>(finding a woman to marry) |

# Work 3: Tracing the toxic meanings

## Representations and identities (social meanings)

- Either beautiful and stupid or ugly and stupid
  - Less intelligent but good at cunning
  - Good at memorizing but less employable
- } Age-old women stereotypes met with modern rhetoric of employment equality
- 
- Virginity complex repackaged as individual freedom
  - Beauty and virginity tradable with  
men's socioeconomic success
  - Victimizing cheating men by complaining them
- } Traditional requests to women repackaged by capitalism and human rights

# Work 3: Tracing the toxic meanings

## Representations and identities (social meanings)

- Blaming women for illegal actions against them  
e.g., asking for physical violence
  - Downplaying the criminality of illegal actions  
e.g., buying women because of poverty
- Blameworthy and expendable victims
- 
- Acceptance of non-virtuous wives
  - Penalizing with stigmatized feminism labels  
e.g., woman boxers, feminism cancer
- Marked gender asymmetry  
Stigmatized gender symmetry



## Next step

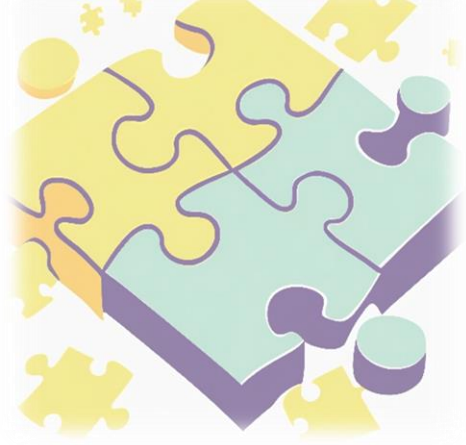
- Based on the identified biased meanings, develop **performance criteria** to test gender biases embodied in AI output
  - Starting with a RA in April (received grant from UURIP program)
  - Applied for OpenAI researcher access support (US\$12,000 API deployment)

## Next step

- Based on the identified biased meanings, develop **performance criteria** to test gender biases embodied in AI output
  - Starting with a RA in April (received grant from UURIP program)
  - Applied for OpenAI researcher access support (US\$12,000 API deployment)



A dataset of (implicit) toxic language produced by AI



## Where meanings meet the toxic language

- **Work 1:** Human ways of connecting meanings to their language
- **Work 2:** AI's capability of associating meanings to its own language or input
- **Work 3:** A dataset of (implicit) biased language produced by AI

# Where meanings meet the toxic language

- **Work 1:** Human ways of connecting meanings to their language
- **Work 2:** AI's capability of associating meanings to its own language
- **Work 3:** A dataset of (implicit) biased language produced by AI



In-context learning paradigms/SFT meaning layer that associates biased language output to toxic meanings and use it to improve LLM performance


# Where meanings meet the toxic language

## **AI across languages and cultures**

- How does the different understanding of toxic language between East and West influence their AI development and AI performance?  
e.g., political correctness in East and West
- *<Toxic language in East Asia>* (Routledge) [Chen & Brown, in progress]
- Comparative studies of toxic language across cultures
- Testing AI across cultures for the differences in their biased output

# Where meanings meet the toxic language

## AI across languages and cultures

- How does the different understanding of toxic language between East and West influence their AI development and AI performance?  
e.g., political correctness in East and West
  - <Toxic language in East Asia> (Routledge) [Chen & Brown, in progress]
  - Comparative studies of toxic language across cultures
  - Testing AI across cultures for the differences in their biased output
- 
- Seeking the common ground in developing ethical AI across languages and cultures

# Working towards ethical AI

Working with  
psycholinguists and  
neurolinguists

- Generalise human-based findings on meaning connections into a framework

Working with  
sociolinguists and  
pragmaticians

- Identifying toxic meanings and their constructions

Working with  
computational  
linguists and NLPers

- Developing in-context learning paradigms and/or SFT pre-trained models for meaning association

# Upcoming work

- Work under review

| Index | Type of publication | Publisher                             | Expected publication | Current status             |
|-------|---------------------|---------------------------------------|----------------------|----------------------------|
| [1]   | Journal article     | <i>Applied Linguistics</i>            | 2024                 | Major correction completed |
| [2]   | Journal article     | <i>Journal of Politeness Research</i> | 2024                 | Major correction completed |
| [3]   | Journal article     | <i>Discourse, Context and Media</i>   | 2025                 | Invited contribution       |
| [4]   | Journal article     | Toxic language about women            | 2025                 | Submitted                  |
| [5]   | Journal article     | Metapragmatics 'modesty' in Chinese   | of 2025              | Manuscript completed       |



# Upcoming work

- Work underway

| <b>Index</b> | <b>Type of publication</b> | <b>Topic</b>   | <b>Expected publication</b> | <b>Current status</b>                   |
|--------------|----------------------------|--|-----------------------------|---|
| [6]          | Journal article            | Gender biases learnt by AI                                   | 2025                        | Received internal grant                 |
| [7]          | Journal article            | Gap between ChatGPT and humans in retrieving meanings        | 2025                        | Data analysis in progress               |
| [8]          | Journal article            | ‘Marginalised’ masculinity                                   | 2025                        | Data analysis completed                 |
| [9]          | Journal article            | Longitudinal development of form-context-meaning connections | 2025/2026                   | Data collected and ethics approved      |
| [10]         | Edited book                | Toxic language in East Asia                                  | 2025/2026                   | First draft expected by the end of 2024 |

Thank you!

# References

- Agha, A. (2005). Registers of Language. In *A Companion to Linguistic Anthropology* (pp. 23–45). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470996522.ch2>
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 29.
- Chang, Yuh-Fang. 2011. 'Interlanguage Pragmatic Development: The Relation between Pragmalinguistic Competence and Sociopragmatic Competence'. *Language Sciences* 33 (5): 786–98. <https://doi.org/10.1016/j.langsci.2011.02.002>.
- Chen, X., & Brown, L. (2022). Second Language knowledge of pragmatic meanings: What do learners of Korean know about the Korean pronouns ce and na? *Journal of Pragmatics*, 202, 7–22. <https://doi.org/10.1016/j.pragma.2022.10.010>
- Chen, X., & Lee, J. (2021). The relationship between stereotypical meaning and contextual meaning of Korean honorifics. *Journal of Pragmatics*, 171, 118–130. <https://doi.org/10.1016/j.pragma.2020.10.011>
- Chen, X., Li, J., & Ye, Y. (2024). A feasibility study for the application of AI-generated conversations in pragmatic analysis. *Journal of Pragmatics*, 223, 14–30. <https://doi.org/10.1016/j.pragma.2024.01.003>
- Cook, H. M. (2011). Are honorifics polite? Uses of referent honorifics in a Japanese committee meeting. *Journal of Pragmatics*, 43(15), 3655–3672. <https://doi.org/10.1016/j.pragma.2011.08.008>
- Deng, J., Zhou, J., Sun, H., Zheng, C., Mi, F., Meng, H., & Huang, M. (2022). COLD: A Benchmark for Chinese Offensive Language Detection. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 11580–11599). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.796>
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In J.R. Firth (Ed.) *Studies in linguistic analysis* (pp.1-32). oxford: Philological society.

# References

- House, Juliane, and Dániel Z. Kádár. 2021. 'Altered Speech Act Indication: A Contrastive Pragmatic Study of English and Chinese Thank and Greet Expressions'. *Lingua* 264 (December): 103162. <https://doi.org/10.1016/j.lingua.2021.103162>.
- Kasper, Gabriele and Kenneth R. Rose. 2001. 'Pragmatics in language teaching'. In Rose, Kenneth R., and Gabriele Kasper. (Eds). *Pragmatics in Language Teaching*. Cambridge: Cambridge University Press. 1-9.
- Lenci, A., & Sahlgren, M. (2023). *Distributional Semantics*. Cambridge University Press. <https://doi.org/10.1017/9780511783692>
- Malik, J. S., Qiao, H., Pang, G., & Hengel, A. van den. (2023). *Deep Learning for Hate Speech Detection: A Comparative Study* (arXiv:2202.09517). arXiv. <https://doi.org/10.48550/arXiv.2202.09517>
- Okamoto, S. (2011). The use and interpretation of addressee honorifics and plain forms in Japanese: Diversity, multiplicity, and ambiguity. *Journal of Pragmatics*, 43(15), 3673–3688. <https://doi.org/10.1016/j.pragma.2011.06.012>
- Roever, Carsten. 2011. 'Testing of Second Language Pragmatics: Past and Future'. *Language Testing* 28 (4): 463–81. <https://doi.org/10.1177/0265532210394633>.
- Taguchi, Naoko. 2006. 'Analysis of Appropriateness in a Speech Act of Request in L2 English'. *Pragmatics* 16 (4): 513–33. <https://doi.org/10.1075/prag.16.4.05tag>.
- Taguchi, Naoko. 2011. 'Do Proficiency and Study-Abroad Experience Affect Speech Act Production? Analysis of Appropriateness, Accuracy, and Fluency' 49 (4): 265–93. <https://doi.org/10.1515/iral.2011.015>.
- Taguchi Naoko. 2015. 'Cross-cultural adaptability and development of speech act production in study abroad'. *International Journal of Applied Linguistics* 25 (3): 343–65. <https://doi.org/10.1111/ijal.12073>.
- van Compernelle, Rémi A. 2014. 'Sociocultural Theory and L2 Instructional Pragmatics'. In *Sociocultural Theory and L2 Instructional Pragmatics*. Multilingual Matters. <https://doi.org/10.21832/9781783091409>.