Integrating talker and message in language processing: the influence of speaker gender on sentence prediction in Mandarin Chinese

Yun Feng (yunfeng@polyu.edu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong

Yao Yao (y.yao@polyu.edu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong

Ming Xiang (mxiang@uchicago.edu)
Department of Linguistics, The University of Chicago, Chicago, USA

Abstract

Spoken sentence processing often requires integrating the linguistic message with information about the talker and the social context. Among other things, information about the talker's gender identity could influence how a listener processes what they hear, due to the prevalence of gender-related stereotypes in human societies. Several previous studies showed that listeners anticipated stereotype-consistent content, and that comprehension was affected when gender stereotypes were violated (e.g., when women serve in conventionally male-dominant professions, or *vice versa*). However, the existing findings are rather mixed. In this study, we examine the influence of talker gender information on language comprehension and prediction in Mandarin Chinese. We report the results of a cloze task where participants were asked to guess the last word of a sentence with or without information about talker gender. When talker gender information was available, we further varied whether gender information was revealed by personal names or voices. Participant responses were evaluated for gender bias by two pre-trained language models based on Word2Vec and GPT2. Results of statistical analysis revealed that participants adjusted their responses to align with the gender category cued by the sentence (i.e., more male/female-biased responses when the sentences implicated a male/female talker), but the effect was only present when gender information was implicated through names but not voices. The current study provides partial evidence for the effect of talker gender on sentence prediction. We discuss the implications of current findings for further study of the integration of linguistic and social information in language comprehension.

Keywords: sentence prediction; gender stereotype; language model; Mandarin Chinese

Introduction

In language comprehension, comprehenders engage not only with the literal linguistic content of sentences but also heavily draw upon various contextual information. This includes the broader linguistic context, general world knowledge, and the interlocutors' social identity, all of which contribute to the comprehension of speaker intentions and sentential meaning. One important question is how a language user's access to social conventions and stereotypes influences their language processing. It has been shown that when the literal linguistic message and contextually accessible social information are incompatible, language comprehension may be challenged, compelling the listener to reassess the intended meaning behind the words spoken. For example, we may all find it strange to hear a child asking for wine in a restaurant. But the specific language processing account remains underspecified.

In this study, we focus on the influence of gender-related conventions and stereotypes that prevail in human societies (in the rest of the paper, we use *stereotypes* as a cover term for both conventions and stereotypes). Gender-related stereotypes project individuals onto specific professions, hobbies, attires, personality traits, *etc*. based on their gender identity. For example, firefighters are often assumed to be male, and nurses female. Prior literature provided some evidence from behavioral and *event-related brain potential* (ERP) data that violations of gender-related stereotypes could affect language processing at both the lexical level and the sentence level, but the results are quite mixed regarding the exact nature of the consequences of such violations (see Porkert et al., 2024 for a review).

At the word level, it has been shown that presenting a word loaded with gender-related stereotypes (e.g., *lipstick*) could prompt a stereotype-consistent gender-marked word (e.g., *women* or *she*) and *vice versa*. Several studies (Pesciarelli et al., 2019; Siyanova-Chanturia et al., 2012; Wang et al., 2017; White et al., 2009) used a lexical priming paradigm where the prime and target words were either stereotype-consistent (e.g., *lipstick–women*) or stereotype-inconsistent (e.g., *lipstick–men*). Overall, these studies found slower and less accurate responses together with a larger N400 effect when the prime and target were gender-mismatched, although Siyanova-Chanturia et al. (2012) only found the N400 effect when a feminine-biased role (*teacher*) was paired with a male pronoun (*he*) but not when a masculine-biased role (*driver*) was paired with a female pronoun (*she*).

At the sentence level, the results are more mixed. Most studies reported a later, P600 effect when the sentence stimuli contained gender stereotype violations. For example, multiple studies (Canal et al., 2015; Irmen et al., 2010; Kreiner et al., 2009; Osterhout et al., 1997; Su et al., 2016) examined the processing of gender-marked referents

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

(pronouns or nouns) with stereotype-imbued antecedents (typically subject nouns) in a reading task, and found a P600 effect—but not an N400 effect—when the referent's gender marking was mismatched with the antecedent's stereotypical gender (e.g., "Our aerobics instructor gave *himself* a break"). Interestingly, the P600 effect was larger for female participants (Osterhout et al., 1997), suggesting that female participants responded more strongly to violations of "appropriate" gender roles. Besides, female subject nouns with male pronouns (e.g., *actress* with *himself*) elicited a larger P600 effect, especially in those participants who described themselves as more feminine (Canal et al., 2015). Meanwhile, a few other studies (Du & Zhang, 2023a, 2023b; Molinaro et al., 2016) reported N400 responses to incongruencies of gender stereotypes in sentence reading tasks. These studies typically presented gender-marked constituents earlier than stereotype-imbued ones in the sentence stimuli (e.g., Li's son is a nurse).

In the auditory domain, Lattner & Friederici (2003) found that when listening to self-referential statements (e.g., "I like to wear lipstick/play soccer") spoken by female/male voices, German listeners showed a P600 response when the voice gender was incongruent with the gender stereotype of the statement. These results contrasted with an earlier study by Van Berkum et al. (2008), which used a similar paradigm with Dutch participants and found an earlier, N400-like response to stereotype-violated stimuli. However, Van Berkum et al.'s stimuli included not only gender-related stereotypes but also stereotypes related to age and social class, which could partially account for the different results.

To summarize, the existing literature suggests that violations of gender stereotypes can be quickly detected in language comprehension. Unlike the classic N400 and P600 effects associated with semantic and syntactic anomalies, the ERP responses elicited by gender-related stereotypical violations are more variable, probably because stereotypical knowledge provides subtle and cancellable cues for sentence prediction and the integration of such cues is influenced by when and how gender information is revealed. According to Porket et al.'s (2024) review, N400 effects would be more likely when gender information is readily available and stereotypical violation can be easily calculated (e.g., in the word priming paradigm or when gender information is revealed earlier than the activation of gender stereotypes), whereas P600 effects would be more likely when reintegration or reassessment is needed, as in anaphoric or self-referential sentence contexts that require stronger inferential processing.

Much of the existing literature is based on languages (English, Dutch, German, Italian, Spanish, etc.) spoken in the western cultural contexts. There are only a few recent studies reporting on Mandarin Chinese (Du & Zhang, 2023a, 2023b; Su et al., 2016; Wang et al., 2016, 2017), although gender stereotypes are inherently cultural-specific and responses to violations of social norms may differ across populations (Mu et al., 2015). Furthermore, the current literature mainly focuses on the effects of stereotypical violation on language comprehension. An implicit assumption is that as the linguistic message unfolds, readers or listeners may generate gender-specific predictions via the activation of gender-related stereotypical knowledge, which further influences subsequent processing. However, to the best of our knowledge, whether language users generate gender-specific predictions based on gender-related cues has not been directly tested.

In the current paper, we report results from a cloze task with Mandarin Chinese speakers. The participants were presented with partial sentence prompts and were asked to complete the sentences with the first one or two words that came to their mind. All the sentence prompts were self-referring (i.e., about 我 "I" or 我的 "my"). We constructed the following conditions by varying whether and how gender cues were present in the sentence fragments: (1) a baseline condition with no gender cues; (2) name-cued conditions where the sentence prompts were preceded by the name of the talker (e.g., 小美说 "Xiaomei says") that was either stereotypically female or male (e.g., 小美 "Xiaomei" is a common female name in Chinese); (3) voice-cued conditions where the sentence prompts were spoken by female or male voices. Our main hypothesis is that participants would generate sentence predictions that are aligned with the gender of the talker of the sentence. Under the baseline condition, participants might consider themselves as the talker of the sentence by default since the sentences were self-referring and there was no additional talker gender information. Under the name-cued and voice-cued conditions, the participants had access to the gender information of the talker (either via names or via talker voices). Thus, we predict that the sentence predictions in the baseline condition would be aligned with the participant's own gender, whereas the predictions in the gender-cued conditions would be aligned with the gender group cued by the names/voices.

Methods

Participants

A total of 210 native Mandarin Chinese speakers participated in the study (Mean age = 24.02, SD = 2.90; 106 F, 104 M). All the participants grew up in Mainland China and use Mandarin as their dominant language. None of them reported any vision, hearing, or neurological disorders. The study was approved by the ethics committee at the institution of the authors. All the participants were compensated for their time with course credit or a small amount of monetary reward.

Materials and Procedures

Materials. The experimental stimuli consisted of 189 self-referring sentence prompts in Chinese (e.g., 我最喜欢的虚拟形象是__ "My favorite virtual image is ___", 我的朋友都

喜欢我的___ “My friends all love my ___”). The prompts were incomplete sentences with the final word missing. The sentence fragment prompts cover a wide range of attributes that any gender group could be associated with, including hobbies, routines, attires, professions, and personality traits, etc; but there may be culture-specific gender-norms when these attributes are discussed for different gender groups.

The sentence prompts were presented in five conditions: a baseline condition with no additional gender cues and four gender-cued conditions that varied by cue type (personal names vs. voices) and cued gender category (female vs. male) in a 2×2 design. In the baseline and name-cued conditions, the stimuli were only presented visually; in the voice-cued conditions, the stimuli were presented both visually and auditorily. With a between-participant design, each participant was only tested on one of the five conditions.

In the name-cued conditions, the personal names (24 female names and 24 male names) were initially selected based on the Chinese national census data (2019 to 2021) and recommendations of ChatGPT as gender-unambiguous names that are commonly used by current young Chinese speakers. The names were further vetted by a separate group of native Mandarin speakers (N = 12; Mean age = 23.92, SD = 2.87; 8 F, 4 M) who rated the names' familiarity (1 = "not familiar"; 5 = "very familiar") and gender association with each gender (1 = "not associated with female/male"; 5 = "highly associated with one female/male") on a 5-point Likert scale. Overall, as shown in Table 1, the names were rated as highly familiar (typically above 4) and strongly associated with the target gender (Mean > 4.65) but not the non-target gender (Mean < 1.49). There was no significant difference between female and male names in terms of familiarity (t = .864, p > .1) or gender bias (t = 1.68, p > .1).

Table 1: Gender associations and familiarity of the Chinese names on a 5-point scale.

	Female name (N = 24)	**Male name** (N = 24)
Familiarity	4.36 ± 0.37	4.27 ± 0.31
Association with female	4.76 ± 0.26	1.49 ± 0.37
Association with male	1.24 ± 0.25	4.65 ± 0.22

The personal names were randomly assigned to precede the sentence prompts in the name-cued conditions (see Table 2 for examples), with each name appearing on 6–8 prompts.

In the voice-cued conditions, the auditory stimuli were created by recording a female talker (mean pitch = 210.6 Hz) and a male talker (mean pitch = 172.4 Hz) reading the sentence prompts in a soundproof booth. The recording was done via Praat with a sampling rate of 44.1 kHz. Both talkers were native Mandarin speakers in their 20s without any noticeable accent; they were instructed to maintain a stable and emotion-neutral intonation (pitch, intensity, speech rate) across sentences.

To reduce fatigue, the sentence prompts were divided into two balanced lists, each with 94-95 stimuli. Each participant only completed one list. Overall, each sentence prompt was evaluated by at least 20 participants (balanced in gender) in each condition.

Procedure. The experiment was administered on the PCIbex platform (Schwarz & Zehr, 2021), and participants completed the study in a computer lab on campus equipped with desktop computers and headsets (for the voice-cued conditions). Participants were instructed to read/listen to the sentence prompts and type from a keyboard the first Chinese phrase (in 1–3 Chinese characters) that came to their mind that would complete the sentence.

Data preprocessing and analysis

Responses from the cloze task were evaluated for gender biases with a computational approach using pre-trained language models. In a nutshell, we followed the methodology of Li et al. (2022) and Nadeem et al. (2020), and calculated gender bias by comparing the semantic similarity between the cloze task responses and a set of pre-defined gender-referential words, with semantic similarity quantified as the cosine similarity between word embeddings. The gender-referential word lists, taken from Li et al. (2022), which was in turn adapted from Nadeem et al. (2020), consisted of eighteen male referential words (e.g., 爸爸 "dad", 男人 "men") and eighteen female referential words (e.g., 妈妈 "mom", 女人 "women"). More details of the calculation of gender bias are presented below.

Table 2: Examples of sentence prompts in each condition.

Condition	**Cue type**	**Cued talker gender**	**Examples of sentence prompts and English translations** (personal names were replaced with common female/male names in English for readers' convenience)
Baseline	NA	NA	我最喜欢的虚拟形象是 __。(My favorite virtual image is __.)
FemaleName	Name	Female	小美说：我最喜欢的虚拟形象是 __。(Mary said: My favorite virtual image is __.)
MaleName	Name	Male	小辉说：我最喜欢的虚拟形象是 __。(Tom said: My favorite virtual image is __.)
FemaleVoice	Voice	Female	🔊 (female voice) 我最喜欢的虚拟形象是 __。(My favorite virtual image is __.)
MaleVoice	Voice	Male	🔊 (male voice) 我最喜欢的虚拟形象是 __。(My favorite virtual image is __.)

For better generalizability, we calculated semantic embeddings with two models, Word2Vec and GPT2, both of which were pre-trained on large, comprehensive Chinese corpora. The Word2Vec embeddings[1] were generated using the Skip-Gram with Negative Sampling (SGNS) method and pre-trained on Baidu Encyclopedia Corpus (4.1 GB, 745M tokens; Li et al., 2018). The GPT2 embeddings [2] were pre-trained on the Chinese CLUECorpusSmall dataset (14 GB, 5 billion characters).

Two sets of gender bias metrics were generated using the two language models, respectively. With each model, average semantic embeddings were calculated for each sentence prompt by averaging across the word embeddings of all the responses generated in the cloze task for that sentence prompt. This was done for responses generated by female participants and male participants separately, under each experimental condition. Next, average semantic embeddings were separately calculated for the list of female and male referential words. We then calculated the cosine similarities between the average semantic embeddings of the human responses (per sentence prompt, condition, and participant gender) and those of the female referential words ($Similarity_{female\_words}$) and male referential words ($Similarity_{male\_words}$). The similarity measures quantified the alignment of human responses with the semantic spaces of female or male-referential words; in this study, we calculated the overall gender bias of the human responses as the difference between the two similarities (GenderBias = $Similarity_{male\_words} - Similarity_{female\_words}$). Thus, a gender bias of zero would indicate that a given sentence prompt has no overall gender bias (i.e., gender-neutral), whereas a positive value would indicate an overall male bias, and a negative value indicates a female bias.

Both sets of gender bias metrics (one derived from Word2Vec and the other GPT2) were modelled for potential effects of experimental condition and participant gender. More specifically, we used two model structures. The first model (Model 1) used the complete dataset and included the effects of condition (treatment-coded; reference = Baseline), participant gender (treatment-coded; reference = Female), and their interaction as fixed effects and SentencePrompt as a random effect (Model 1 formula: GenderBias ~ Condition * ParticipantGender + (1 | SentencePrompt)). The goal of Model 1 was to compare the Baseline condition with the other four gender-cued conditions (FemaleName, MaleName, FemaleVoice, MaleVoice) in terms of the effect of participant gender. We predict that participant gender will have a significant effect on responses in the Baseline condition, when participants used their own genders to predict sentence completions, but the effect should be reduced in the gender-cued conditions, when gender cues of an external talker were available.

The second model (Model 2) used a partial dataset that only included the gender-cued conditions and excluded data from the baseline condition. Model 2 included cued talker gender (treatment-coded; reference = Female), participant gender (treatment-coded; reference = Female), cue type (treatment-coded; reference = Name) and their interactions as fixed effects and SentencePrompt as a random effect (Model 2 formula: GenderBias ~ CuedTalkerGender * ParticipantGender * CueType + (1 | SentencePrompt)). The goal of Model 2 was to evaluate the effects of cued talker gender across conditions. We predict that cued talker gender would have a significant effect across the board and that the size of the cued talker gender effect might vary with participant gender and cue type.

All the models were built with the lmer() function in the LmerTest package (Kuznetsova et al., 2017) in R (R Core Team, 2021). Nonsignificant ($p > .05$) higher-order interactions were trimmed, and only the final models are presented here.

Results

Data overview

The cloze task yielded 19875 responses (3386 types) with an average length of 2.28 Chinese characters (SD = 0.71). About 5% of the responses contained words not existent in the language models (mainly proper names) and were thus excluded from the analysis.

Table 3: Top five female- and male-biased words in the cloze responses according to Word2Vec and GPT2.

	Word2Vec	GPT2
Female-biased	口红 “lipstick” 漂亮 “pretty” 苗条 “slim” 内衣 “underwear” 花瓶 “vase”	裙子 “dress” 妈妈 “mom” 拖鞋 “slipper” 短袖 “T-shirt” 短裤 “shorts”
Male-biased	成龙 “Jackie Chan” 霍金 “Hawking” 军人 “soldier” 帅哥 “handsome guy” 才华 “talent”	爱因斯 “Einstein” 军人 “soldier” 钢笔 “pen” 宇航 “astronaut” 聪明 “smart”

Both language models generated reasonable gender bias metrics ($M_{Word2Vec}$ = 0.0098, SD = 0.018; M_{GPT2} = –0.0013, SD = 0.0054). As shown in Table 3, the five most female-biased words in the cloze responses according to the language models were mainly related to feminine appearances (e.g., 漂亮 “beautiful”), makeup and attires (口红 “lipstick”), whereas the most male-biased words were related to competence (e.g., 才华 “talent”) and male-dominant careers (e.g., 军人 “soldier”). In a separate study, we tested the consistency between model-generated gender bias metrics

[1] https://github.com/Embedding/Chinese-Word-Vectors

[2] https://huggingface.co/uer/gpt2-large-chinese-cluecorpussmall

and human ratings with a separate set of 192 Chinese words, and found significant positive correlations for both Word2Vec (r = .80, p < .001) and GPT2 (r = .58, p < .001).

While there is some cross-model variation in terms of the exact values of gender bias scores, gender bias metrics from both models show very similar patterns of variations across experimental conditions. As shown in Figure 1, in the baseline condition, as expected, the cloze responses aligned with participants' own gender, with female participants giving more female-biased responses and male participants giving more male-biased responses. What is surprising, however, is that the FemaleVoice and MaleVoice conditions patterned similarly with the baseline condition, showing no effect of the voice cue. On the other hand, in the FemaleName and MaleName conditions, the effects of participants gender were largely overridden and the responses seemed to be aligned with the perceived gender of the talker, with female talker names eliciting more female-biased responses and male talker names eliciting more male-biased responses.

Mixed-effects models, based on either Word2Vec or GPT2-based word embeddings, confirmed these observations (Table 4). Model 1 revealed a significant effect of participant gender for the baseline condition (Word2Vec: β = 0.009, t = 7.1, p < .001; GPT2: β = 0.002, t = 5.3, p < .001), indicating that male participants gave more male-biased response than female participants in the baseline condition. Compared to the baseline condition, the effect of participant gender was significantly reduced in the FemaleName (Word2Vec: β = –0.008, t = –4.0, p < .001; GPT2: β = –0.001, t = –2.4, p = .02) and MaleName (Word2Vec: β = –0.007, t = –3.3, p < .001; GPT2: β= –0.002, t = –3.4, p < .001) conditions, but no significant difference between baseline and voice-cued conditions was detected regarding the effect of participant gender (all ps > .1). In other words, the pattern of participants giving responses in line with their own gender was present in the baseline and voice-cued conditions but largely disappeared in the name-cued conditions.

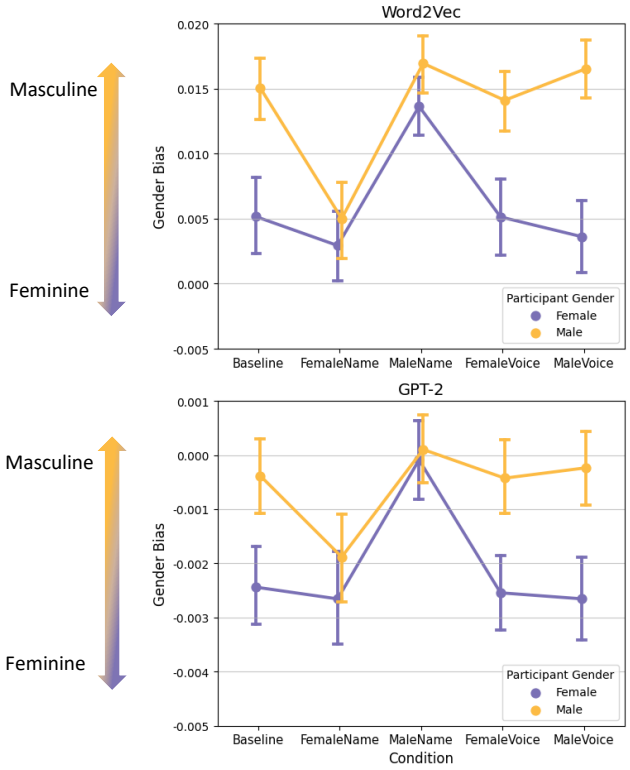

Figure 1: Mean (with standard errors) gender biases across conditions generated by Word2Vec and GPT2 embeddings.

Table 4: Summary of fixed effects in Models 1 and 2 based on Gender Bias metrics from Word2Vec and GPT2

		Word2Vec			GPT2		
	term	***β***	***t***	***p***	***β***	***t***	***p***
Model 1	(intercept)	0.005	3.9	< 0.001	-0.002	-6.4	< 0.001
	PtpGender = Male	0.009	7.1	< 0.001	0.002	5.3	< 0.001
	Cond=FemaleName	-0.002	-1.6	0.1	-0.0002	-0.6	0.6
	Cond=MaleName	0.008	6.1	< 0.001	0.002	6.1	< 0.001
	Cond=FemaleVoice	-0.00002	-0.02	0.9	-0.0001	-0.3	0.8
	Cond=MaleVoice	-0.002	-1.1	0.3	-0.0002	-0.6	0.5
	PtpGender=Male:Cond=FemaleName	-0.008	-4.0	< 0.001	-0.001	-2.4	0.02
	PtpGender=Male:Cond=MaleName	-0.007	-3.3	< 0.001	-0.002	-3.4	< 0.001
	PtpGender=Male:Cond=FemaleVoice	-0.0009	-0.5	0.6	0.00006	0.1	0.9
	PtpGender=Male:Cond=MaleVoice	0.003	1.5	0.1	0.0004	0.7	0.5
Model 2	(intercept)	0.003	2.6	0.009	-0.003	-6.8	< 0.001
	CuedTkrGender = Male	0.010	8.3	< 0.001	0.002	6.9	< 0.001
	PtpGender = Male	0.001	1.2	0.25	0.0005	1.6	0.1
	CueType=Voice	0.002	1.3	0.21	-0.0001	-0.4	0.7
	CuedTkrGender=Male:PtpGender=Male	0.003	1.9	0.06	-0.0001	-0.4	0.7
	CuedTkrGender=Male:CueType=Voice	-0.01	-7.8	< 0.001	-0.002	-5.7	< 0.001
	PtpGender=Male:CueType=Voice	0.008	5.9	< 0.001	0.002	4.6	< 0.001

On the other hand, when only data from gender-cued conditions were modelled (i.e., Model 2), significant effects of cued talker gender were found in the name-cued conditions (Word2Vec: $\beta = 0.010$, $t = 8.3$, $p < .001$; GPT2: $\beta = 0.002$, $t = 6.9$, $p < .001$), indicating that male talker names elicited more male-biased response than female names did. But no effect of participant gender was found in the name-cued conditions (ps > .1), suggesting that responses in the name-cued conditions were **not** affected by the participant's own gender. Compared to the name-cued conditions, the effect of cued talker gender was greatly reduced in the voice-cued conditions (Word2Vec: $\beta = -0.010$, $t = -7.8$, $p < .001$; GPT2: $\beta = -0.002$, $t = -5.7$, $p < .001$) while the effect of participant gender rose to significance (Word2Vec: $\beta = 0.008$, $t = 5.9$, $p < .001$; GPT2: $\beta = 0.002$, $t = 4.6$, $p < .001$).

Taken together, the modeling results indicate that the responses in the name-cued conditions (FemaleName and MaleName) were mainly aligned with the cued talker's gender in terms of gender bias, but the responses in the voice-cued conditions (FemaleVoice and MaleVoice) were mainly aligned with the participant's gender, similar to the baseline condition when there was no information about the talker.

Discussion

In this study, we examine the hypothesis that comprehenders' word prediction in sentence processing would be influenced by their social stereotype knowledge based on the talker's gender. We conducted a series of sentence completion tasks (i.e., cloze task) in Mandarin while controlling for participant gender and varying whether and how talker gender was revealed. Analysis of the gender biases in participants' responses—as estimated by two pre-trained language models—provided partial evidence for the hypothesis. Specifically, in the baseline condition, when there was no other talker in the context, the responses were aligned with participants' own gender, consistent with theories of language prediction that call upon one's own production model (Pickering & Gambi, 2018; Pickering & Garrod, 2013). By contrast, in the name-cued conditions, when talker gender was revealed by conventionally gendered personal names, participant responses were more stereotypically aligned with the talker's cued gender and largely dissociated from the participants' own gender. However, in the voice-cued conditions, when talker gender was revealed through voice features, participant responses were mainly aligned with their own gender—similar to the baseline condition—as if the participants were ignoring the talker voices.

Thus, we observe unambiguous evidence of the participants taking the perspective of the contextually cued talker and almost overriding their own perspective in the name-cued conditions. These results indicate that it is possible for language users to predict another talker's speech by taking into consideration the social identity of the other talker and stereotypical knowledge about the social group the other talker belongs to. These findings are consistent with previous ERP results showing that violations of gender stereotypes could be detected during sentence reading tasks.

However, we did not observe any reliable effects of talker gender in the voice-cued conditions, which seems to contradict previous studies that found effects of gender stereotype violation in spoken sentence comprehension (Grant et al., 2020; Lattner & Friederici, 2003; Van Berkum et al., 2008). We offer several possible explanations. First of all, it is possible that although participants both heard and read the sentence prompts, they somehow paid more attention to the visual stimuli, which were self-referring sentence frames in written forms, and therefore generated the sentence completions from their own first-person perspectives. A related explanation is that some participants might think that the completed sentences would be attributed to themselves—although they received instructions to "complete the sentences for the talker" before the experiment began—and therefore tended to use their own perspectives. In the post-hoc inspection of the data, we indeed found that some participants provided sentence completions that were most likely from their own instead of the target talker's perspectives (e.g., 我喜欢夏天穿裙子 "In the summer, I like to wear dresses" submitted by female participants for a voice-cued trial with a male voice).

Secondly, while the personal names in the name-cued conditions were vetted for familiarity and gender bias, the voices used in the voice-cued conditions did not go through rigorous testing. Although none of the participants had difficulty identifying the voice genders, the voices may not be sufficiently typical to activate certain stereotypes, which raises a question of which voice characteristics are more likely to trigger gender stereotypes. Along the same line, the fact that there were only two voices (1 F, 1 M) and that each participant only heard one voice throughout the session would also increases the likelihood for participants to get habituated with and pay less attention to the voice cues over the course of the experiment. By comparison, 24 personal names were used for each gender in the name-cued conditions, and the variability of names may have contributed to the participants' continuing attention to the name cues.

In this study, gender biases were evaluated by comparing the word embedding similarities between the participants' responses and a list of pre-defined gender-referential words. This method has the potential to scale up to large-sample investigations of other kinds of social stereotypes in language use, but its validity also needs to be firmly established. We plan to conduct follow-up studies to collect human ratings on the gender biases of the response words and compare with the word embedding-based estimates.

Another limitation of the current study is that we only made use of two gender categories, female and male. This obviously is an over-simplification. People's perception of gender categories is much more nuanced and goes beyond a binary distinction. Future studies need to take this into account and adopt a more realistic experimental design.

Acknowledgments

This project was funded by research grants from the Hong Kong Polytechnic University (No. P0046383 and No. P0051039). We thank Shiyue Li for her help with data collection.

References

Canal, P., Garnham, A., & Oakhill, J. (2015). Beyond Gender Stereotypes in Language Comprehension: Self Sex-Role Descriptions Affect the Brain's Potentials Associated with Agreement Processing. *Frontiers in Psychology*, *6*.

Du, Y., & Zhang, Y. (2023a). Discourse Context Immediately Overrides Gender Stereotypes during Discourse Reading: Evidence from ERPs. *Brain Sciences*, *13*(3), Article 3.

Du, Y., & Zhang, Y. (2023b). Strategic Processing of Gender Stereotypes in Sentence Comprehension: An ERP Study. *Brain Sciences*, *13*(4), Article 4.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to Me: The Social Context of Infant-Directed Speech and Its Effects on Early Language Acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344.

Grant, A., Grey, S., & van Hell, J. G. (2020). Male fashionistas and female football fans: Gender stereotypes affect neurophysiological correlates of semantic processing during speech comprehension. *Journal of Neurolinguistics*, *53*, 100876.

Irmen, L., Holt, D. V., & Weisbrod, M. (2010). Effects of role typicality on processing person information in German: Evidence from an ERP study. *Brain Research*, *1353*, 133–144.

Kemper, S. (1994). Elderspeak: Speech accommodations to older adults. *Aging, Neuropsychology, and Cognition*, *1*(1), 17–28.

Kreiner, H., Mohr, S., Kessler, K., & Garrod, S. (2009). *Can context affect gender processing? ERP differences between definitional and stereotypical gender* (K. Alter, M. Horne, M. Lindgren, & J. von Koss Torkildsen, Eds.; pp. 107–119). Lund Universitet.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26.

Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing – evidence from event-related brain potentials. *Neuroscience Letters*, *339*(3), 191–194.

Li, J., Zhu, S., Liu, Y., & Liu, P. (2022). Analysis of Gender Bias in Social Perception and Judgement Using Chinese Word Embeddings. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 8–16.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). *Analogical Reasoning on Chinese Morphological and Semantic Relations* (arXiv:1805.06504). arXiv.

Molinaro, N., Su, J.-J., & Carreiras, M. (2016). Stereotypes override grammar: Social knowledge in sentence comprehension. *Brain and Language*, *155–156*, 36–43.

Mu, Y., Kitayama, S., Han, S., & Gelfand, M. J. (2015). How culture gets embrained: Cultural differences in event-related potentials of social norm violations. *Proceedings of the National Academy of Sciences*, *112*(50), 15348–15353.

Nadeem, M., Bethke, A., & Reddy, S. (2020). *StereoSet: Measuring stereotypical bias in pretrained language models* (arXiv:2004.09456). arXiv.

Osterhout, L., Bersick, M., & Mclaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*(3), 273–285.

Pesciarelli, F., Scorolli, C., & Cacciari, C. (2019). Neural correlates of the implicit processing of grammatical and stereotypical gender violations: A masked and unmasked priming study. *Biological Psychology*, *146*, 107714.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

Porkert, J., Siyanova-Chanturia, A., Loerts, H., Schüppert, A., & Keijzer, M. (2024). N400 or P600?—A Systematic Review of ERP Studies on Gender Stereotype Violations. *Language and Linguistics Compass*, *18*(5), e12530.

R Core Team, R. S. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

Schwarz, F., & Zehr, J. (2021). Tutorial: Introduction to PCIbex – An Open-Science Platform for Online Experiments: Design, Data-Collection and Code-Sharing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Siyanova-Chanturia, A., Pesciarelli, F., & Cacciari, C. (2012). The Electrophysiological Underpinnings of Processing Gender Stereotypes in Language. *PLOS ONE*, *7*(12), e48712.

Su, J.-J., Molinaro, N., Gillon-Dowens, M., Tsai, P.-S., Wu, D. H., & Carreiras, M. (2016). When "He" Can Also Be "She": An ERP Study of Reflexive Pronoun Resolution in Written Mandarin Chinese. *Frontiers in Psychology*, *7*.

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(4), 580–591.

Wang, P., Yang, Y.-P., Tan, C.-H., Chen, Q.-W., & Cantfort, T. (2017). Gender Stereotype Activation versus Lexical Semantic Activation: An ERP Study. *The Journal of General Psychology*, *144*(4), 283–308.

Wang, P., Yang, Y.-P., Tan, C.-H., Zhao, X.-X., Liu, Y.-H., & Lin, C.-D. (2016). Stereotype activation is unintentional: Behavioural and event-related potenials evidence. *International Journal of Psychology*, *51*(2), 156–162.

White, K. R., Crites, S. L., Jr, Taylor, J. H., & Corral, G. (2009). Wait, what? Assessing stereotype incongruities

using the N400 ERP component. *Social Cognitive and Affective Neuroscience*, *4*(2), 191–198.