A feasibility study for the application of AI-generated conversations in pragmatic analysis

Xi Chen, Jun Li, Yuting Ye

Abstract

This study explores the potential of including AI-generated language in pragmatic analysis – a field that has primarily been conducted on human language use. With the rapid growth of large language models and high-performing chatbots, AI-generated texts and AI-human interactions constitute a growing field where pragmatics research is expanding to. Language data that humans used to hold a full authorship may also involve modifications made by AI (e.g., AI proofreading). The foremost concern is thus the pragmatic qualities of AI-generated language, such as whether and to which extent AI data mirror the pragmatic patterns that we have found in human speech behaviours. In this study, we compare 148 ChatGPT-generated conversations with 82 human-written ones as well as 354 human evaluations of these conversations. The data are analysed using various methods, including traditional speech strategy coding, four computational methods developed in NLP, and four statistical tests. The findings reveal that ChatGPT performs equally well as human participants in four out of the five tested pragmalinguistic features and five out of six sociopragmatic features. Additionally, the conversations generated by ChatGPT exhibit higher syntactic diversity and a greater sense of formality compared to those written by humans. As a result, our participants are unable to distinguish ChatGPT-generated conversations from human-written ones.

Keywords: ChatGPT; speech act; pragmalinguistic; sociopragmatic; pragmatic competence

1. Introduction

Recent advancements in large language models (LLMs) and chatbots, such as ChatGPT, have surprised their users with how well AI produces coherent, relevant, and even appropriate texts. At the same time, they have raised concerns about the potential infiltration of AI-generated messages in various domains, such as news reports, school coursework, and legal documents. In addition, AI-generated texts and AI-human interactions are becoming a new and growing field where pragmatic issues need to be examined. A surge of studies has compared AI-generated content with that produced by human participants, including student essays (Herbold et al. 2023), abstracts of academic papers (Ma et al. 2023), and medical texts (Liao et al. 2023). A few psycholinguistic studies have also examined linguistic hypotheses to assess the extent to which language choices made by AI resemble those of humans (Cai et al., 2023; Qiu et al, 2023). However, except Qiu et al (2023) who found a possible deficiency of ChatGPT in processing pragmatic inference, the pragmatic performance/competence of AI has not yet been examined systematically.

The current study investigates the pragmalinguistic and sociopragmatic competence of AI by comparing 148 AI-generated and 82 human-written conversations that are elicited using

74 speech act scenarios. It tests five pragmalinguistic features and six sociopragmatic features of the conversations, using a variety of analytical methods, including traditional qualitative analysis used in pragmatics, four computational techniques developed in NLP (Natural Language Processing), and four statistical tests. The study aims to provide a comprehensive account for the feasibility of including AI-generated language as a data source in pragmatic analysis, which has thus far been conducted on human language use.

The feasibility study holds significant implications for pragmatics as a scientific discipline and language education where the acquisition of pragmatic competence plays an essential role. As mentioned at the beginning of this paper, there are, at least, two new types of AI data that become available for pragmatic analysis. One is AI-generated texts and AI-human interactions where AI holds an authorship for the language it outputs. Another is human speeches and texts that are modified by AI. However, before incorporating them into research agenda of pragmatics, discussions are needed regarding their pragmatic qualities and the pragmatic competence of AI. This study makes one of the first few steps in providing a comprehensive analysis of AI-generated conversations. Albeit not without limitation in the types of AI data it uses (i.e., textual conversations), its findings provide an experimental ground for future pragmatics studies to investigate AI performance in different contexts.

In the context of language education, AI, especially conversational AI, is increasingly used as a collaborative partner to human teachers (Ji, Han and Ko, 2023). Conversations generated by it provide language educators with teaching and assessment materials as well as prevent educators from repetitive practices that consume a large amount of class time. Language learners also receive real-time feedback from chatbots. They provide language learners access to language learning materials and complement the insufficient opportunities that they have for authentic communication. Moreover, previous studies find that conversing with AI can reduce the anxiety of L2 learners when communicating in their L2 (Shao et al, 2019). Again, one fundamental concern that underlies all the claimed benefits of integrating AI and its data into language learning is the pragmatic qualities of AI-generated conversations. In other words, the conversations generated by AI need to be examined for their human-like quality before AI's assistance is embraced in both pragmatics and language education.

Below, we begin by reviewing extant tests of pragmatic competence, with light shed on both pragmalinguistic and sociopragmatic features. We then move on to our methodological design, data collection and data analysis (Section 3). Findings are presented in Section 4 including both quantitative and qualitative results. They are followed by a discussion and a conclusion on what AI-generated conversations can offer pragmatics studies and what is still in the 'black box'.

2. Testing pragmatic competence

2.1. Pragmalinguistic competence and sociopragmaic competence

Pragmatic competence can be traced back to Hymes' (1966) proposal of communicative competence and, as one of its components, pragmatic competence overlaps arguably with

sociolinguistic competence (Bachman, 1990; van EK, 1986). Albeit having some variations in its definitions, previous studies largely agree that pragmatic competence is the ability to use language to deliver the speaker's intention, convey and interpret meanings beyond literal meanings, and achieve specific actional purposes (Fraser, 2010; Ishihara and Cohen, 2010; Thomas 1983). Purpura (2004) divides it into knowledge, ability, and performance. And, indeed, many studies have focused on one or two of these components, for example, pragmatic awareness (e.g., Bardovi-Harlig and Griffin 2005; Cheng 2016), sociopragmatic knowledge (e.g., Chen and Ren, 2023), pragmatic ability (e.g., Cohen 1996), and pragmatic performance (e.g., Bella 2011). However, there is no clear distinction between these elements. Many of the studies have examined one by investigating the other, for example, examining pragmatic competence from observing pragmatic performance. Just as Laughlin, Wain, and Schmidgall (2015, 6) noted, "performance is competence that can be observed".

Leech (1983) contributed one of the most popular categorizations of pragmatic competence/performance, namely, pragmalinguistic competence/performance and sociopragmatic competence/performance. The former refers to "the resources for conveying communicative acts and relational or interpersonal meanings" and the latter examines "the social perceptions underlying participants' interpretation and performance of communicative action" (Kasper and Rose, 2011,2). For example, at an interview, interviewees carefully choose their language, avoid the use of slang, shape their expressions, and organize their speeches. Their ability to manipulate these language resources pertains to pragmalinguistic competence. Their control of language resources is guided and accompanied by their awareness of the interview context and their perceptions of what kind of language should be used in this context (e.g., formality, politeness). Such awareness is addressed as their sociopragmatic competence. The two types of competences are closely interrelated with the sociopragmatic competence directing pragmalinguistic choices and the pragmalinguistic competence affecting the realization of sociopragmatic perceptions. Nevertheless, they are often assessed by different factors and in different ways.

With the focus on language resources, pragmalinguistic competence has been assessed based on grammatical accuracy and discourse control (Taguchi 2006), semantic mitigation moves and clarity of illocutionary force (Taguchi 2015), diversity of speech strategies (Chang 2011), and conventional expressions (House and Kadar 2021). The commonly used analytical approach is to identify and/or categorise pragmalinguistic features into different patterns and calculate the frequency of each pattern. For example, when refusing, speakers may express their 'inability', make 'excuses', and offer an 'alternative' (see, for example, Beebe et al., 1990). These speech strategies often appear in different frequencies, showing the diversity of language choices made by the speakers.

In contrast, sociopragmatic competence has often been examined using human judgments, including participants' evaluations of directness, politeness, appropriateness and formality (Taguchi 2006; Taguchi 2011), their understanding of situational variables (Chang 2011; van Compernolle 2014), and their choice of adhering to or flouting social norms. The participant-oriented approach was a shift from the traditional approach that mapped sociopragmatic concepts (e.g., politeness) to speech strategies, e.g., using honorifics was

regarded as a polite strategy (Blum-Kulka and Olshtain 1984; Brown and Levinson 1987). After the rise of first-order politeness, more attention was paid to the evaluative nature of these sociopragmatic concepts, namely, whether a speech strategy is polite, indirect, or formal should be evaluated based on the speaker's intention and the hearer's interpretation (Chen and Wang, 2021; Eelen 2001; Mills and Grainger 2016). In L2 pragmatics studies, it is not rare to see that appropriateness and adherence to social norms are evaluated (Chen, 2022; Economidou-Kogetsidis 2016). We should emphasize that the divorce between pragmalinguistic forms and sociopragmatic concepts has not torn the two types of pragmatic competence apart, but instead created space for the flexible relationships that exist between them.

2.2. Testing pragmatic competence via speech acts

Speech acts have been one of the most popular avenues for assessing both pragmalinguistic and sociopragmatic competence. Developed by Austin (1962), speech acts connect one's language use to its performative functions, such as using language to request, refuse, or complain. The famous CCSARP (Cross-Cultural Speech Act Realization Project) in the 1980s led to a blossom of analysing speech acts by categorizing their speech strategies – an important feature of pragmalinguistic competence (Blum-Kulka and Olshtain, 1984, Blum-Kulka, 1987, Blum-Kulka et al,1989). For example, request speech acts were divided into head acts and peripheral moves, each of which consisted of various speech strategies. 'Query preparatory' (e.g., would you…) and 'grounders' (e.g., offering a reason) were found to be the predominant request strategies in English (Economidou-Kogetsidis 2013; Fukushima 1996). 'Reasons', 'regrets' and 'gratitude' were found frequent in English refusals (Shishavan and Sharifian 2013; Takahashi and Beebe 1987).

Recent studies additionally explored the possibility of using other pragmalinguistic features than speech strategies to understand the complexity of speech act performance. Su (2017) and Su and Fu (2023) employed local grammar to analyse the function-grammatical patterns of speech acts in English and Chinese. They provided a fine-grained phraseological analysis, although local grammars have limitations in illustrating the relations between different discourse units, i.e., discourse relations, and are affected by the word order that different languages have (Su and Fu 2023). House and Kadar (2021) were interested in the corresponding relationship between recurrent expressions (namely, conventionalised expressions) and speech acts, such as *hello* and the act of greeting. By calculating the frequencies of conventionalised expressions and their contextual variations, such as *thank you* used with complaints instead of gratitude, House and Kadar found that Chinese and English were different in the types of speech acts that the same thanking expressions may convey.

Many of these pragmalinguistic features are now associated with sociopragmatic evaluations, in contrast to the traditional form-concept mappings. Taguchi (2006), for example, considered the influence that ungrammatical utterances and less organized discourses may have on participants' rating of appropriateness. She incorporated the examination of grammaticality and discourse organization into participants' evaluations of appropriateness

of request performance. For example, point 5, which was assigned to the highest level of appropriateness, was described as "expressions are fully appropriate for the situation. No or almost no grammatical and discourse errors" (2006:520). While this approach highlighted the interconnections between pragmalinguistic features and sociopragmatic evaluations, it provides little specification on the actual syntactic and discoursal properties of a conversation, such as the diversity of syntactic constructions that participants adopted and their ways of organising discourses. The current study adopts computational methods to address such limitations, including calculations of syntactic diversity and identification of discoursal relations.

We should also note that although the aforementioned studies have suggested multiple pragmalinguistic and sociopragmatic features, these features have scarcely been assessed altogether due to the different research purposes that previous studies had. In addition, they have rarely, if not never, been applied to assess the pragmatic performance of AI. In this study, we make one of the first steps to analyse all the aforementioned features with AI-generated data, including five pragmalinguistic ones, i.e., lexical, syntactic, strategic, discoursal, and conventionalised features, and six sociopragmatic features ranging from understanding of contexts to adherence to social norms. In doing so, we assess ChatGPT-generated conversations using the same metrics that have been used to test various aspects of the pragmatic competence of human participants.

3. Methodology

We collected 82 conversations from human participants and 148 from ChatGPT as well as 354 human evaluations of these conversations. Strategy coding and computational techniques, such as NLTK, were employed to analyse pragmalinguistic features (see details in Table 1), and statistical tests were conducted to examine sociopragmatic features (see details in Table 2).

3.1. Participants

A total of 42 participants were recruited on a voluntary basis. They were students at a U.K university, studying different subjects including language studies, marketing, and forensic studies at both undergraduate and postgraduate levels. Four of them were aged between 18 and 19, 36 between 20 and 29, and two between 30 and 39. There were 33 females, four males, three non-binary, and two others. We were not oblivious of the influence that imbalanced gender and age ratio may have on speech act performance. In this study, we took the potential influence into consideration when designing experiments and reporting data analysis results. The mitigating measures include (1) choosing prompts that were close to the participants' life experience. One of the given roles in the prompt was often designed to be their age peer, for example, a person who has 'classmates' or needs to ask for an extension for coursework; (2)

minimising gender indicators in the prompts to allow participants to envisage the gender of interlocutors at their preference; (3) reporting the results by taking into consideration previous findings on language use by different age and gender groups, for example, younger generation's writing was found to be influenced by their extensive use of short messages (Rosen et al. 2010). However, we should note that there were often not agreed conclusions but rather contradictive findings on the gender influence on language use. For example, Fatemeh, Naji and Abdulah (2018) found that female and male English speakers differed significantly in their use of refusal speech strategies while Nelson, Batal and Bakary (2002) argued that they did not.

Among the participants, there were 38 L1 English speakers and one L1 speaker each for Czech, Portuguese, Polish, and Hungarian. The four non-L1 English speakers had an advanced level of proficiency in English which allowed them to study in an English-speaking programme. 37 of the 38 L1 English speakers had at least one second language with varied proficiency levels. Only one of the L1 English speakers was monolingual. The current study intentionally maintained the non-L1 and multilingual English speakers who provided data of 'world English'. Similarly, ChatGPT was trained on large-scale datasets obtained from internet, books, websites, and other texts that unlikely consisted only of L1 English, although OpenAI has not publicly disclosed the specifics of the individual datasets.

3.2. Experimental design

Our experiment design consisted of two parts: the design of effective prompts for eliciting conversations and the design of a questionnaire for collecting emic sociopragmatic evaluations.

We built our prompts on the scenarios that had previously been used to elicit speech act performance, whether they were part of DCTs (Discourse Completion Tasks) or role plays. We excluded the scenarios that were duplicated in different studies and collected a total of 212 different scenarios from 36 academic papers that investigated English speakers and were published between 1984 and 2022 (Appendix A). We started our prompt design with these examined scenarios because they had been proven effective in generating dialogic data from English-speaking participants, at least.

74 scenarios (Appendix B) were selected from the 212, based on the following criteria: (i) the scenario presented a situation that was close to the participants' life experience, e.g., scenarios related to campus life. This selection helped our participants to produce more real-like conversations, on the one hand. On the other, it provided ChatGPT with a role that was similar to the age group of the participants, and hence minimised the influence of imbalanced age distribution; (ii) the list of scenarios contained a variety of speech acts, including 20 for eliciting requests, 13 for refusals, 9 for complaints, 8 for apologies, 7 for suggestions, 5 for compliment response, 4 for advice; 2 for compliments, 2 for gratitude, 2 for invitations, 1 for offer and 1 for regret. The different number for each speech act was a result of the disproportionate studies on them; (iii) the scenarios were varied by contextual variables.

Speakers in 21 of them had less power than their hearers, 41 with equal power, and 12 with greater power. Similarly, speakers were distant in 21 of the scenarios, acquaintance in 33, and intimate in 20. By including a diversity of different speech acts and scenarios, we aimed to ensure that the data collected were not biased by the types of speech acts or by the types of role relationships. Data analysis results were thus more generalisable across different contexts.

Before applying the selected scenarios to collect data, five student research assistants (RAs) re-wrote them by (i) minimising gender indicators in them, for example, "a man who came to the shop" was rewritten as "a person who came to the shop". The gender-neutral descriptions allowed our participants of different genders to envisage the gender of their interlocutor at their preference; (ii) setting the speakers as "you" and someone (e.g., your friend, professor, a stranger) to prevent ChatGPT from making a third-person casted conversation. The RAs also modified the format of these scenarios by testing the rewritten versions with ChatGPT. This was to find an effective prompt template that generated conversations consistently in the AI, instead of narratives or monologues. We had regular meetings to discuss different templates and finalised them as Example (1) shows:

Example (1) Situation 9

Use a maximum of 6 sentences ("turns" in the version to human participants) to make a dialogue for the situation below:

A friend of a classmate called and asked to borrow some class notes of yours. You had agreed to meet him/her at the library that afternoon, but you forgot about it. That evening she calls explaining she waited for an hour for you.

Write the conversation as if you feel apologetic and as if the classmate's friend feels annoyed.

The template started with a clear instruction "Use a maximum of 6 sentences …". "Sentences" were used here instead of 'turns' because ChatGPT recognised 'sentence' as the whole utterance that one speaker took his/her turn for. In the version that we offered our participants, "sentences" were changed back to "turns". Meanwhile, the number of sentences was defined to a maximum of six because ChatGPT tended to expand a conversation when there was not a clear limit and eventually drifted away from the given scenario. Participants were allowed to write shorter or longer than the given limit, although most of their conversations remained around six turns. The second part of the template was a description of the scenario. It consisted of two characters "you" and another person whose relationship with "you" was specified. The incident between the two characters was also described clearly. The template ended with another instruction that provided one or two attitudinal indicators in accordance with the context, for example, one would typically feel 'apologetic' if s/he missed an appointment. The attitudinal instruction was found effective in eliciting natural conversations from ChatGPT possibly because the AI was evidenced to have a human-like neuropsychological ability to identify emotions correctly (Loconte et al. 2023). It was neither

uncommon to spot an attitudinal description in traditional speech-act elicitation scenarios (e.g., "You are annoyed by the loud music next door"). Only, for the ease of AI's understanding, we made it explicit in a separate command line.

RAs rewrote the 74 scenarios, again, using the template. The formatted scenarios were then assigned to both ChatGPT and human participants to collect conversations, with a slight modification between 'sentences' and 'turns'.

In the meantime, we designed a sociopragmatic questionnaire to collect participants' interpretations of six sociopragmatic features that previous studies had suggested, namely, understanding of context, appropriateness of strategy use, levels of politeness, levels of (in)directness, proper-ness of formality, and the extent to which social norms were adhered to or flouted. Example 2 below showcases the value assignment to different questionnaire items. Details of the questionnaire can be found in Appendix C.

Example (2)

Q2. In the conversation, did the speakers use appropriate strategies to communicate with each other?

5 Strategy use was fully appropriate.

4 Strategy use was mostly appropriate.

3 Strategy use was somewhat appropriate.

2 Strategy use was largely INappropriate

1 Strategy use was entirely INappropriate.

Q3. Did the conversation have a proper level of politeness?

3 Yes. Proper level of politeness

2 More polite than I would expect

1 Less polite than I would expect

This approach to the sociopragmatic features was in line with the constructivist approach to pragmatics, namely, the evaluation of each sociopragmatic feature was based on participants' emic judgment (Chen and Wang, 2021; Eelen 2001; Mills and Grainger 2016). The item in the questionnaire started with a question, requesting the participants to judge the behaviour of both the 'speakers' or the 'conversation'. It intentionally led the participants to pay attention to both interlocutors of a conversation and their communicative exchange, considering that appropriateness and politeness are not a result of one-way intention but rather achieved in reciprocal exchanges (Culpeper and Tantucci, 2021; Tantucci et al., 2022).

The question was followed by level descriptions of the sociopragmatic feature in a 5-point Likert scale. The descriptions (e.g., fully appropriate, mostly appropriate) followed the design used in previous studies (Cunningham 2017; Taguchi 2006). These studies have examined and evidenced their clarity and effectiveness in rating speech act performance. Only, politeness and (in)directness were tested against the participants' expectations, as their nature is highly subjective to the participants' ideological beliefs (Chen and Wang, 2021; Eelen 2001). For example, being 'fully polite' with a friend may not be appropriate because the level of politeness might have diverged from the friend's expectation. At the end of the questionnaire, we also included an additional item asking the participants to discern whether a conversation was made by AI or by humans.

3.3. Data collection

Data were collected in two steps. In Step 1, we collected conversations written by human participants and generated by ChatGPT, using the formatted 74 scenarios. In Step 2, we collected human participants' emic evaluations of sociopragmatic features for each of the collected conversations using the designed sociopragmatic questionnaire. The two steps led to two types of data: textual conversations and numeric ratings.

In Step 1, each human participant was required to write two conversations, one for each of the two scenarios that had been randomly assigned to them. 37 of the 42 participants completed this step with two of them enthusiastically writing a few more, resulting in a total of 82 conversations collected.

At the same time, ChatGPT was used by the RAs to generate two conversations for each scenario. RAs were required to open a 'new chat' when generating each conversation. This measure was to prevent ChatGPT from learning from its previous productions. A total of 148 conversations were collected from the AI.

In Step 2, one human-written conversation was juxtaposed in random order with two ChatGPT-generated conversations in the same scenario. They were aligned in terms of format and attached with sociopragmatic questionnaires. The three-in-one bundle was sent to the human participants to collect their evaluations of six sociopragmatic features and their discernment between AI and human conversations. Each participant was assigned nine conversations (three scenarios, each with three conversations). We should note that none of the participants was assigned the conversation that they wrote. In other words, they were all rating the conversations that were either written by other participants or by the AI. We asked them not to compare between conversations, but to focus on evaluating the properties of each conversation. They were only informed of that there might be AI-generated ones, but were not acknowledged how many and where these AI-generated conversations were placed. These measures are intended to prevent participants from feeling obligated to find AI-generated conversations in the bundle of three conversations.

38 of the recruited participants completed Step 2, including three who had not completed Step 1. These three participants volunteered to rate a few more conversations to

compensate for their absence in Step 1. In total, 207 conversations were rated, with 147 conversations in 49 of the 74 scenarios being rated repeatedly by, at least, two participants. Another 60 in 20 scenarios were rated once. Unfortunately, we did not receive ratings on 15 conversations in 5 scenarios.

We should note that the numbers, i.e., one conversation from participants in each scenario and two from ChatGPT as well as nine for each participant to evaluate, are the result of a careful balance between the participants' workload and data size that the current study needs. In other words, we intentionally maintained a manageable amount of work for participants to ensure that they were not overwhelmed by the task. At the same time, we secured sufficient data for the following analysis.

3.4. Data analysis

Data analysis began by calculating the inter-rater reliability, which examined the agreements that participants had between them in terms of sociopragmatic understanding. We then conducted the analysis of pragmalinguistic features and sociopragmatic features.

Wilcoxon Signed-Rank test based on Difference-in-Difference (DiD) (Abadie, 2005; Bertrand et al, 2004) was employed to evaluate the rating consistency across multiple raters on the same scenarios. Compared to Cohen's kappa which was traditionally used in linguistics research, our testing scheme based on DiD works for more than two raters, takes into account their individual biases, such as the tendency to assign high/low scores, and is easier for interpretation. The results displayed a non-significant p-value for each of the rating items (p=0.9891), indicating the study is established on an acceptable level of inter-rater agreement.

We then adopted both traditional strategy coding and computational methods to analyse five pragmalinguistic features that previous studies had repeatedly suggested (Blum-Kulka and Olshtain 1984; Chang 2011; Taguchi 2006). Table 1 presents each of the pragmalinguistic features and the methods that were used to explore them.

Table 1. Tested pragmalinguistic features

Pragmalinguistic features	**Methods**	**Specifications**
Language choice	Lexical diversity	Lexical diversity was calculated using NLTK (Natural Language Toolkit). It divided the total number of words used in one conversation by the number of different words used in that conversation.
	Syntactic diversity	Syntactic diversity was similarly calculated based on the number of

		different dependency trees divided by the total number of dependency trees in each conversation, using NLTK.Tree.
	Discourse relation	Using Java End-to-End Discourse Parser developed based on PDTB (Penn Discourse Treebank) style, we extracted explicit discourse relations that occurred five times or more.
Conventional expressions	Weighted average and quantile of frequency	Weighted average of occurrence per 100 files was calculated for the most frequent 15 expressions used by ChatGPT and participants, separately, in each of the included speech acts. Frequency quantile was then calculated for each identified expression *across* all included speech acts (except offer and regret which did not have enough participant data). This method examined whether an expression is conventionalised to a specific speech act or simply prevalently used in all speech acts. For instance, 0.75 indicates that an expression is more frequent in one speech act than 75% of the other speech acts.
Strategy use	Strategy categorization	We coded speech strategies for requests and refusals, using the established coding schemes (Beebe et al., 1990; Blum-Kulka and Olshtain 1984; Su and Ren 2017). These two speech acts were coded because they had the most scenarios, generated the most conversations, and had widely accepted coding schemes.

Specifically, we employed computational methods to provide a quantitative overview of the language choices made by the AI and human participants at three levels: lexical, syntactic, and discoursal. Syntactic diversity concerns the ability of a speaker to achieve a pragmatic purpose or meet the pragmatic constraint by using a variety of syntactic

constructions (Bardovi-Harlig, 2012; Bybee, 2010; Delin et al., 1996; Li et al., 2023). Discourse relations refer to the connective relationships between elementary discourse units, for example, 'cause' when two units were reasons and results and 'contrast' when one unit rejected the other. We then zoomed in onto specific speech acts to compare the conventional expressions that were anchored to them (House and Kadar 2021) and the strategies that ChatGPT and human participants used to realize them.

Conventional expressions were initially related to indirect speech acts, referring to linguistic means that do not explicitly mark a speech act but can be recognised so by social conventions (Blum-Kulka and Olshtain 1984; Blum-Kulka 1987). It is defined as having three components: recurrent sequences, context-dependence, and social contract (Bardovi-Harlig, 2012). The 'sequences' often consist of multiple linguistic units (Edmond, 2014). Accordingly, we examined all possible combinations across 2 to 5-grams and extracted 15 expressions that occurred the most frequently in each n-gram. We then manually extracted and combined the expressions that were repeatedly included across 2 to 5-grams, for example, 'thank you' in 2-gram, 'thank you for' in 3-gram, 'thank you so much' in 4-gram, and 'thank you for inviting me' in 5-gram were subsumed to 'thank you'. In this process, we also calculated those unigrams that frequently occurred in multiple grams, such as 'please'. The subsumption follows Bybee's argument that conventionalization is more often applied to meaning than being bound to fixed forms (see, 2010, pp.151, 153, 157). With respect to the different contexts which conventional expressions occur, we conducted the extraction based on the type of speech acts. Thus, requests and refusals were found to have different conventional expressions (Table 5 in Section 4.1.2). We also took a step further to identify the degree to which a recurrent sequence was contracted to a speech act, using quantile ranking. It examines whether the expression occurs more frequently in performing one specific type of speech act than the others. If an expression occurs more frequently in other speech acts, it may be contracted to express the illocutionary point of the other speech act, for example, 'thank you' is frequent in requests but more contracted to gratitude, or prevalently used in different speech acts without an 'anchored' value. In carrying out the above three calculations (n-gram, frequency, quantile ranking), we examined the three defined characteristics of conventional expressions. We should also emphasize that this study is not set to extract an exhaustive list of conventional expressions for specific speech acts. Instead, it focuses on comparing AI and participants in terms of their choice of conventional expressions.

Sociopragmatic features were examined based on human participants' ratings of each conversation using the designed sociopragmatic questionnaire. Table 2 listed the sociopragmatic features, their rating system, and the statistical tests used to compare them.

Table 2. Tested sociopragmatic features

Sociopragmatic features	**Rating system**	**Statistical tests**
Understanding of contexts	5-point scalar ratings between "understood very well" and "did not understand at all"	

Appropriateness of strategy use	5-point scalar ratings between "fully appropriate" and "entirely inappropriate"	(paired) permutation test (paired) Wilcoxon signed-rank test Mann-Whitney U test
Level of politeness	3-point scalar ratings between "proper level of politeness", "more polite" and "less polite" than expected	
Level of indirectness	3-point scalar ratings between " proper level of directness", more indirect", and "more direct" than expected	
Proper-ness of formality	5-point scalar ratings between "proper level of formality" and "very improper level of formality"	
Adherence to social norms	5-point scalar ratings between "fully follows the norms" and "flouts seriously the norms"	
Discernment of AI-generated conversations	Binary choice between "AI" and "human"	Chi-square test

The reason that we chose to conduct multiple statistical tests on the same features is to comprehensively investigate the discrepancy between the AI-generated conversations and human-written conversations. In detail, the permutation test (a non-parametric test that relies on fewer assumptions on the score distribution) evaluates the mean difference of the rating scores; the Mann-Whitney U test (used for independent samples) and Wilcoxon signed-rank test (used for dependent samples) look at the median difference, which is robust and insensitive to outlier scores; Chi-square test looks into the distribution difference, which checks if two categorical distributions are the same and thus is the most stringent test (Morgan and Winship, 2007; Rosenbaum, 2002). These tests were implemented in two ways: "paired test" refers to the scores (one for an AI-generated conversation and one for a human-written conversation) by the same reviewer are paired up to eliminate this reviewer's individual bias, similar to the difference-in-difference approach; "independent test" ("independent" omitted if not specified) refers to that the scores for human-written conversations and AI-generated conversations are respectively pooled without considering the connection of score sources.

4. Findings

In this section, we present the comparison results on five pragmalinguistic features first (Section 4.1) and then six sociopragmatic features (Section 4.2). We also provide an answer to the question of whether human participants can distinguish AI-generated conversations from human-written ones (Section 4.3).

4.1. Comparison of pragmalinguistic competence

4.1.1. Diversity of language choice

We explored the diversity of language choice at three different levels, namely, lexical diversity, syntactic diversity, and the diversity of discourse relations. The test results obtained from the computational analysis have shown that ChatGPT and human participants do not differ significantly in terms of the diversity of their lexical choices but differ significantly in their syntactic diversity (Table 3).

Table 3. Paired tests for lexical and syntactic diversity

	Wilcoxon signed-rank test
Syntactic diversity	0.0000
Lexical diversity	0.4191

Specifically, ChatGPT-generated conversations used a greater range of different syntactic structures than the participants, and its performance tended to be consistent. Human participants, on the other hand, appeared to have more individual variations when choosing syntactic structures. Some preferred to diversify their sentence structures while others were conservative, resulting in their level of syntactic diversity being lower than ChatGPT's and their box (individual variation) being wider (Figure 1).

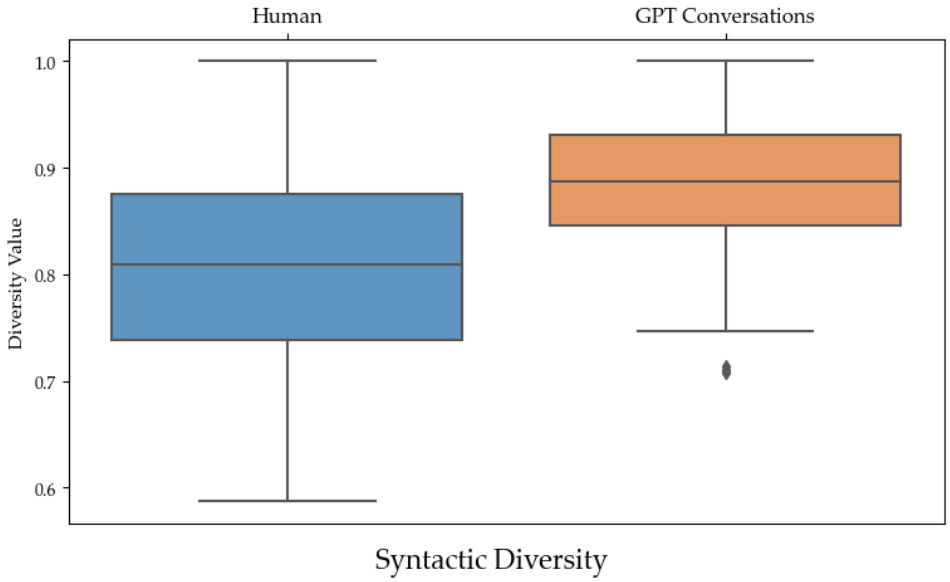


Figure 1. Difference in syntactic diversity

In terms of discourse organization, ChatGPT and participants adopted the same seven types of discourse relations from 30 discourse relations that were provided by PDTB (Penn Discourse Treebank) – the largest discourse relation tree bank (Table 4). We can see that their choices of discourse relations are not restricted, but rather spread across all the four primary categories provided by the PDTB (i.e., comparison, contingency, expansion, and temporal). However, both ChatGPT and participants chose to use the same one or two specific discourse structures in each category. There are some proportional differences, for example, ChatGPT uses more conjunctions and contrasts, while participants employ the others more frequently. However, the broad patterns in their preference are very similar, that is, they both employed contrastive, conditional, and conjunctive structures the most.

Table 4. Proportions of discourse relations

Discourse relation	**GPT**	**Human**
Comparison		
- **Contrast**	0.3433	0.2727
Contingency		
- **Cause**	0.0466	0.1212
- **Condition**	0.1586	0.1602
Expansion		
- **Alternative**	0.0187	0.0216
- **Conjunction**	0.3582	0.2424
Temporal		
- **Asynchronous**	0.0336	0.1082
- **Synchrony**	0.0373	0.0649

4.1.2. Conventional expressions

We calculated the weighted average and quantile of frequency to decide the conventionality of an expression. As detailed in Section 3.4, these methods were applied to identify the most frequent expressions that occur most likely with a specific speech act. The test results do not provide an exhaustive list of conventional expressions, but rather serve to compare ChatGPT and participants in terms of their choice of conventional expressions. Here, we report requests and refusals only. Their datasets were larger than others (63 conversations for requests and 39 for refusals), and hence produced more generalisable patterns (Table 5).

Table 5. Conventional expressions used by ChatGPT and participants in requests and refusals

Request		Refusal	
GPT	Human	GPT	Human
please	please	but i	but i

62.5 (1.00) can/could you 52.5 (1.00) i want(ed) to 35.0 (1.00) (I) really appreciate 30.0 (1.00) i was wondering 22.5 (1.00) get back to 15.0 (1.00) look forward to 7.5 (1.00) thank you 112.5 (0.80) make sure 37.5 (0.80) apologize 30.0 (0.70) talk to 20.0 (0.70)	52.38 (1.00) can/could you 42.86 (1.00) if i 38.1 (1.00) be able to 19.05 (1.00) i am trying to 4.76 (1.00) i was hoping/wondering 33.33 (0.90) you have time 4.76 (0.90) of course 33.33 (0.80) sorry 76.19 (0.70) thank you 71.43 (0.70) (really) want(ed) to 42.86 (0.50)	150.0 (1.00) i understand 100.0 (1.00) i appreciate 65.38 (1.00) how about 38.46 (1.00) i hope 34.62 (1.00) allow me to 23.08 (1.00) that works for 19.23 (1.00) want(ed) to 57.69 (0.90) is there any way 11.54 (0.90) it would be 23.08 (0.80) sorry 42.31 (0.70) i apologize 26.92 (0.70) thank you 53.85 (0.30)	100.0 (1.00) i could 91.67 (1.00) sorry 108.33 (0.80) want(ed) to 58.33 (0.70) thank you 41.67 (0.40)

*Weighted average is given outside parenthesis and quantile in parenthesis.

Comparing the lists in Table 5 shows that ChatGPT and participants have made considerably similar choices of conventional expressions. In requests, both their lists start from 'please' and 'can/could you', which were frequently used in approximately half of their conversations. These two expressions were more frequent in requests than in the other speech acts included in this study (quantile ranking 1.00), showing their high conventionality to the realization of the request speech act. There are other frequent expressions, such as 'thank you' which was used at least once in 70% of the conversations. However, it was used more frequently in three other speech acts than in requests (quantile ranking: .80 in GPT and .70 in humans). Therefore, its conventionality to request is relatively limited. This finding corresponds to our understanding of 'thank you', that is, its illocutionary point is more contracted to express gratitude than a request.

In refusals, both ChatGPT and participants chose the expression 'but I' the most. It was used in almost every refusal conversation (weighted average: 150 and 100 per 100 documents). A closer examination shows that this structure is often used to connect other frequent expressions, for example, connecting expressions of understanding, appreciation, or apologies

to excuses or expressions of inability (Example 3). At times, it is also used immediately after an affirmative 'Yeah' (i.e., Yeah, but I…). It shifts the speaker's stance of alignment (e.g., I understand, I appreciate, I wish) to his/her violation of the hearer's expectation and hence delivers the intention of refusal. Even when the part after 'but I' was left empty, e.g., I understand but I…, the hearer still recognises the speaker's difficulty in accepting his/her proposal. This finding is supported by van Dijk's (1979) findings on 'but' as a pragmatic connective to protest a request.

Example 3.

a. I understand what you mean, **but I** think my Spanish is already on a level high enough that taking another Spanish course would just be a waste. (from participants)

b. I appreciate the suggestion, Professor Johnson, **but I**'ve already taken Spanish courses in the past and feel confident in my proficiency. (from AI)

c. I wish I could stay and help **but I** have an appointment (from participants)

d. I understand the importance of the task, **but I** have a prior commitment that I can't cancel. (from AI)

e. I am really sorry, **but I** don't think I'll have it done in time. (from participants)

f. I'm really sorry, **but I** won't be able to join you today. (from AI)

Following the 'but I', expressions of understanding and appreciation were used by ChatGPT as if they were conventionalized to refusal speech act. However, human data revealed their context dependence. They were much less found in conversations written by participants. This AI-human difference was further confirmed by our manual coding of speech strategies in the next section.

4.1.3. Speech strategy use

The first author and one RA manually coded request sequences and refusal sequences using established strategy coding schemes (Beebe et al., 1990; Blum-Kulka and Olshtain 1984; Su and Ren 2017). By sequences, we refer to the entire interactive process for making a request or a refusal, including not only head acts, but also its supportive moves and adjuncts, such as greeting and thanking (Su and Ren, 2017). These components could be located in different turns which were also taken into consideration.

Figure 2 demonstrates the proportions of different request strategies adopted by ChatGPT and participants. A high level of similarities in both their strategy choice and strategy distribution is obvious. They have chosen an almost identical range of strategies, and the cross-group differences are marginal. In fact, only a 0.75% proportional difference is found

in the most frequently used head act, query preparatory (e.g., could you…), 3.16% in grounders (e.g., offering a reason), and 0.39% in thanking. The popularity of these strategies supports our findings in Section 4.1.2, namely, conventional expressions for requests feature 'please' and 'can/could you' (the main formula of query preparatory) in both ChatGPT-generated and human-written conversations.

In less used strategies, ChatGPT generated more sweeteners (e.g., that is very kind of you) while participants adopted more confirmation (e.g., yeah, ok). Sweeteners are defined as the mitigative move for downgrading the imposition caused by requests (Blum-Kulka and Olshtain, 1984) and confirmations acknowledge the requestee's recipient of information, showing a status of 'I hear you'. The latter was found to be a strategy for maintaining the reciprocity of information transmission in human interactions (Tantucci and Wang, 2022). To confirm whether the difference arises from different performances of AI and participants, we manually checked how they were used. While the difference in sweeteners seems to stand for the different performance of AI and participants, the use of confirmations is found partly affected by individual participants' repetitive use of this strategy. As Example (4) shows, the participant started four of his/her turns with a "yes/yeah" (lines 3, 7, 9, 11).

Example 4.

1	Student:	Hello Professor, how are you?
2	Professor:	I'm good, I wasn't expecting to see you until our class next Wednesday.
3	Student:	**Yes**, I was actually wanting to speak to you about that. You see, my brother is actually getting married next week.
4	Professor:	That's great to hear, congratulations.
5	Student:	Thank you, it's actually happening on Wednesday.
6	Professor:	During our class?
7	Student:	**Yes**, the ceremony actually begins at 1:00, the time our class usually starts. I know this is last minute, but I wanted to ask if it would be okay for me to miss our class to attend the wedding.
8	Professor:	It is a very important lesson, we will be doing exam preparation. Plus it really is last minute.
9	Student:	**Yeah**, I'm sorry but because of my dad's job, we've had to plan everything quite last minute. It's a small wedding you see.
10	Professor:	Will you be able to do the practice exams? I won't be able to hold any catch-up sessions so you'll have to email them to me.
11	Student:	**Yes**, definitely I'll get them done by Thursday morning latest.
12	Professor:	Okay, that's fine then.

After excluding this idiosyncratic case, the AI-human difference in using confirmation markers reduces to approximately 1.3%.

Figure 3 displays a similar tendency in the refusal strategies used by ChatGPT and participants. The two both offer 'reasons' most frequently when refusing, with very similar

proportions (30% versus 27.6%). However, ChatGPT expresses approximately 5% more unwillingness or inability, empathy, and gratitude, whereas the participants state approximately 4% more regrets (e.g., apologies). This finding supports our observations in conventional expressions (Section 4.1.2), namely, ChatGPT-generated refusals feature more gratitude and expressions of understanding than human-written ones. A careful manual examination finds that ChatGPT repeatedly used expressions of empathy in two conversations and expressions of understanding in four. In particular, the four conversations contributed 10 of 17 instances of gratitude, in contrast to participants who expressed gratitude once in the conversations. It is unclear whether ChatGPT tends to be overly understanding and grateful. However, both the strategies used by AI and participants are perceived as appropriate, as the next section will show.

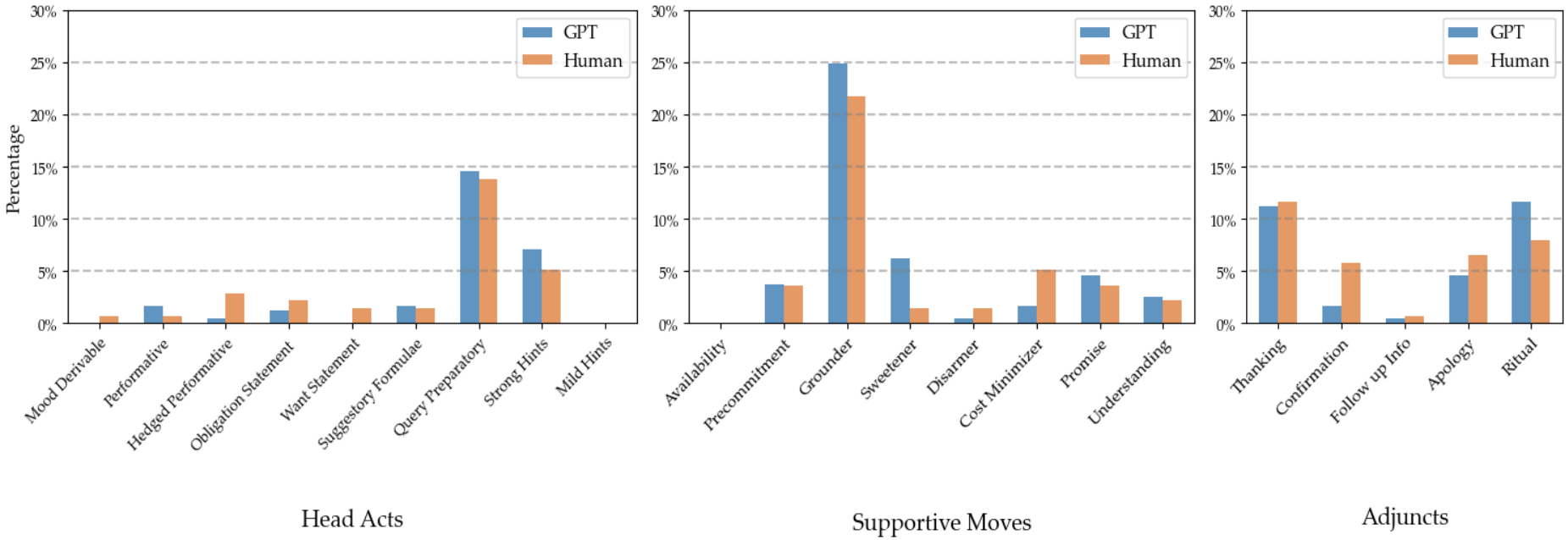
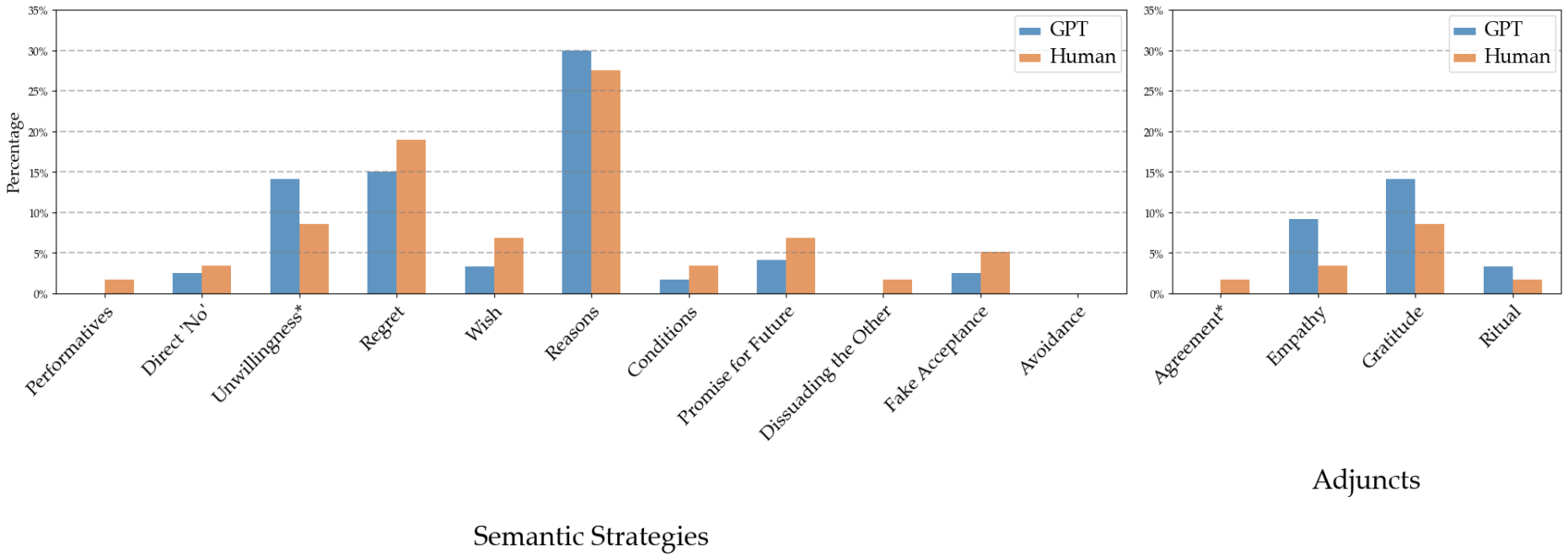


Figure 2. Request strategy use



Refusal

*The theme 'unwillingness' also contains 'inability' which was omitted due to formatting needs. Similarly, the theme 'agreement' also includes 'expressions of positive feeling'.

Figure 3. Refusal strategy use

4.2. Comparison of sociopragmatic competence

We conducted pooled paired tests and pooled independent tests to compare emic evaluations of sociopragmatic features in AI-generated and human-written conversations. The former approach paired up scores, given by the same reviewer, of both AI-generated and human-written conversations, partially mitigating individual biases in assigning higher or lower scores. However, the limited number of pairs might raise doubts about the results. To address this, the latter approach directly aggregated all scores of AI-generated and human-written conversations, respectively. Their findings showed strong agreement and mutually supported each other (Table 6).

Table 6. Tests of sociopragmatic features

Features	**Wilcoxon signed-rank test (paired tests)**	**Mann-Whitney U test (independent tests)**	**Permutation Test (independent tests)**
Context understanding	0.5000	0.6494	0.5350
Appropriateness	0.3095	0.7408	0.4529
Politeness	0.0379	0.8278	0.2316
Indirectness	0.0919	0.3083	0.2918
Proper-ness of formality	0.0001*	0.0006*	0.0009*
Adherence to social norms	0.1648	0.1984	0.3804

According to the test results, ChatGPT and participants do not differ significantly in five of the six sociopragmatic features, including appropriateness of strategy use, politeness, (in)directness, and the extent to which social norms are conformed. The only significant difference is identified with the proper-ness of formality. That is, ChatGPT-generated conversations are rated as more proper in terms of formality than human-written conversations. As Figure 4 clearly exhibits, ChatGPT-generated conversations gained more votes on having the "proper level of formality" (value 5) in contrast to human-written conversations having more with an "acceptable level of formality" (value 3). In other words, ChatGPT has outperformed human participants in choosing the proper level of formality, while performing equally well as humans in the other five sociopragmatic parameters.

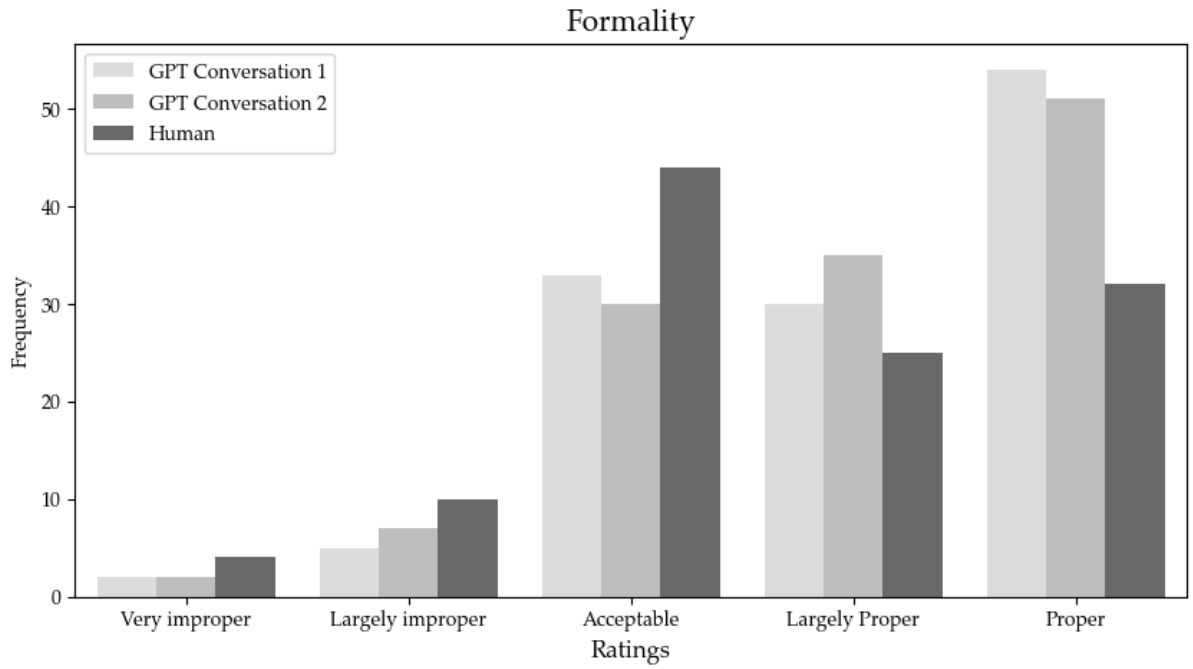


Figure 4. Evaluation of the proper-ness of formality

4.3. Discernment between AI-generated and human-written conversations

The outperformance of ChatGPT seems to have caused difficulties for the participants to discern between AI-generated conversations and human-written ones. They are fairly accurate in identifying human-written conversations, but not so when dealing with ChatGPT-generated conversations. As Table 7 illustrates, the participants have wrongly categorised over half of the conversations, most of which are ChatGPT-generated ones.

Table 7. Frequencies of participants' decisions

Decision	Human-written	GPT conversation 1	GPT conversation 2
Human	93	56	69
GPT	22	70	57

We should reiterate that three conversations in one scenario (one human-written and two GPT-generated) were provided to the participants together in random order. Although the participants were asked not to compare them, the format of having three conversations in a 'bundle' inevitably provided them some advantages to cross-check the conversational contents. However, even with this advantage, the participants still failed to accurately recognize ChatGPT-generated conversations. Chi-square test results show that the participants' accurate decisions on human-written conversations are above chance ($p < 0.0000$), but are no better than chance when identifying ChatGPT-generated conversations ($p > 0.2$). In

brief, human participants are unable to discern ChatGPT-generated conversations even if a human-written reference conversation is provided.

5. Discussion

In this study, we conducted a series of different tests to compare ChatGPT-generated conversations to those conversations written by human participants, with the aim to examine the feasibility of using AI-generated conversations in pragmatic analysis. The findings showed that ChatGPT performed equally well as human participants in four out of the five tested pragmalinguistic features and in five of the six sociopragmatic features. The conversations that it generated also outperformed those of human participants in syntactic diversity and proper-ness of formality. Its high performance resulted in the participants being unable to distinguish ChatGPT-generated conversations from human-written ones.

In terms of pragmalinguistic competence, ChatGPT used a highly similar, at times even identical, range of speech strategies and discourse relations to realize speech acts as human participants. Their patterns in the strategy choice also agree with previous findings. For example, query preparatory and grounders, which we find frequently in both ChatGPT- and human-performed requests, were previously identified as the predominant strategy in British English requests (Fukushima 1996). Especially, query preparatory is the most popular strategy (head act) regardless the requests are written or naturally occurred (Economidou-Kogetsidis 2013). Similarly, the frequent strategies that we identified in refusal speech acts, i.e., 'reasons', 'regret' and 'gratitude', are also found frequently with Anglo-Australian English speakers (Shishavan and Sharifian 2016) and American English speakers (Takahashi and Beebe 1987).

There are some proportional differences in the use of speech strategies and discourse relations between ChatGPT and participants, although the differences tend to be marginal. A possible reason for the small differences is the imbalanced gender ratio that we had in our participants. Previous studies have argued that females tend to feel more apologetic than males (Lazare, 2005). Thus, having more female participants may have resulted in more regret strategies identified with the participants than with the AI. However, contrary to previous findings that female English speakers used more gratitude expressions than their male counterparts (Fatemeh et al., 2018), having more females in our participants did not lead to more gratitude used by humans than by AI. In fact, it was the AI who repetitively used expressions of gratitude. This indicates that gender may not be the only reason for AI-human differences. ChatGPT and participants may also have different levels of sensitivity to contextual variables. For example, sensitivity to different refusal stimuli, namely, refusals to requests, suggestions, invitations or offers, may entail different use of apologies/regret and gratitude (Kwon 2004). These different strategy choices were, nevertheless, perceived as appropriate in sociopragmatic tests.

A third reason for the AI-human difference might lie in the participants' and AI's response bias. That is, if being tested repeatedly, even the same group of participants or the

same AI may give different answers in every round of experiment and would not arrive at the exact same proportions of strategy use in their performance. In fact, it is precisely the variations in strategy use (and other language choices) that indicate speaker identity and stances (Hoshi 2022; Kinginger and Farrell 2004). What is unclear is whether the probabilistic models that LLMs are built upon can effectively convey such sociolinguistic information. Our findings thus call for a further examination on the sociolinguistic competence of ChatGPT and the extent of agreements that ChatGPT and human participants may reach after excluding the random variation in each round of testing.

For testing the sociopragmatic competence, we employed participants' emic perspectives, corresponding to the evaluative nature of the sociopragmatic features (Chen and Wang 2021; Eelen 2001). Of the six tested features, the proper-ness of formality is the only one which has a significant difference with ChatGPT outperforming human participants. This can be attributed, at least partially, to the young university-level participants involved in this study. They are the net-generation whose daily use of texts and short messages is found to affect their ability to write formally (Rosen et al. 2010). The finding also coincides with our impression that university students are currently using AI to rephrase their writings in a 'formal' way.

Besides formality, ChatGPT performed equally well as human participants in understanding contexts, employing appropriate strategies, choosing proper levels of politeness and (in)directness, and adhering to social norms. These findings, together with pragmalinguistic ones, suggest a strong possibility of considering AI-generated conversations as human-like data in pragmatic analysis. The speech act-based AI conversations offer a potential of studying sociopragmatic concepts, such as politeness, as pragmatic effects in addition to human participants' perceptions of them. Moreover, speech act studies, that have often been restricted by sample size and population characteristics (e.g., young, female), may benefit from using AI-generated conversations as a complement to dilute the sample bias. In second language pragmatics, the role of baseline provider has often been played by L1 speakers who are not a homogenous group as criticized (Cook 1999). L2 learners have ideological reasons to resist or reinterpret L1-oriented norms or behaviours (Chen 2022; Chen and Brown, 2022; Ishihara 2008). In this regard, an AI-generated baseline can provide another choice for L2 learners to compare their performance with while evoking less aversion.

For future investigations into AI-generated texts and AI-human interactions, the findings demonstrated that AI could perform pragmatically in a human-like way, given proper conversational prompts. However, prompt design plays a critical role in extracting such performances (Giray, 2023). Our design, namely, instruction + scenario description + attitudinal instruction, represents one of the effective and reliable ways that future studies can start with to obtain AI-generated data. There is not a one-prompt-fits-all. Re-evaluations of human-resemblance might be needed when AI data are generated by other types of prompts. The relationship between human prompts and AI pragmatics also remains to be a task of future studies that are interested in AI-human interactions.

6. Conclusion

To conclude, AI-generated conversations have an equal or even better pragmalinguistic and sociopragmatic performance than human participants. They can be used in pragmatic analysis as a data source on their own and as a complement to human language data, for example, using them to build a baseline that is less biased by speakers' demographic characteristics and subjectivity. However, whether AI-generated conversations can convey effectively information of speakers' subjectivity, such as their identities and stances, still remains as a question. Furthermore, it is unclear how ChatGPT deploys its language resources to be properly 'formal', 'polite', and 'direct'. Human participants rely on their reflexive ability (i.e., metapragmatic ability) and "habitual and instinctive knowledge" that they develop over their language socialisation from a young age (Gumperz 1982, p.162) to do so. The different mechanism that LLMs have from humans raises a concern as to the metapragmatic ability of AI in comparison to human participants, which we encourage future research to address.

We should also provide some caveats regarding the limitations of ChatGPT-generated conversations. First, different genres of AI language, such as reports and professional writtings, need to be examined in terms of their pragmatic qualities, in addition to the genre of conversation that this study investigated. Second, ChatGPT offered only written texts by the time of the current data collection, which may be insufficient in representing the paralinguistic means (e.g., hesitation, pitch contour) used in oral interactions. In this direction, multimodal pragmatic analysis of naturally occurred oral interactions may in turn inform the development of LLMs and help AI to 'speak' in human-like ways. Lastly, ChatGPT does not perform equally well in other languages as it does in English. For research interests in other languages, selecting an LLM and its Chatbots that have been specifically trained using the target language data would be the first step.

References

Abadie, Alberto. 2005. 'Semiparametric Difference-in-Differences Estimators'. *The Review of Economic Studies* 72 (1): 1–19. https://doi.org/10.1111/0034-6527.00321.

Austin, John Langshaw. 1962. *How to Do Things with Words*. Oxford University Press.

Bardovi-Harlig, Kathleen. 2012. 'Formulas, Routines, and Conventional Expressions in Pragmatics Research'. *Annual Review of Applied Linguistics* 32 (March): 206–27. https://doi.org/10.1017/S0267190512000086.

Bardovi-Harlig, Kathleen, and Robert Griffin. 2005. 'L2 Pragmatic Awareness: Evidence from the ESL Classroom'. *System* 33 (3): 401–15. https://doi.org/10.1016/j.system.2005.06.004.

Beebe, Leslie, Tomoko Takahashi, and Robin Uliss-Weltz. 1990. 'Pragmatic transfer in ESL refusals.' In Scarcella, Robin, Elaine Anderson and Stephen Krashen (eds.),

Developing Communication Competence in a Second Language. 55-73. New York: Newbury House.

Bella, Spyridoula. 2011. 'Mitigation and Politeness in Greek Invitation Refusals: Effects of Length of Residence in the Target Community and Intensity of Interaction on Non-Native Speakers' Performance'. *Journal of Pragmatics*, Postcolonial pragmatics, 43 (6): 1718–40. https://doi.org/10.1016/j.pragma.2010.11.005.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. 'How Much Should We Trust Differences-in-Differences Estimates?' *The Quarterly Journal of Economics* 119 (1): 249–75.

Blum-Kulka, Shoshana. 1987. 'Indirectness and Politeness in Requests: Same or Different?' *Journal of Pragmatics* 11 (2): 131–46. https://doi.org/10.1016/0378-2166(87)90192-5.

Blum-Kulka, Shoshana, and Elite Olshtain. 1984. 'Requests and Apologies: A Cross-Cultural Study of Speech Act Realization Patterns (CCSARP)1'. *Applied Linguistics* 5 (3): 196–213. https://doi.org/10.1093/applin/5.3.196.

Blum-Kulka, S., House, J., & Kasper, and G. 1989. 'Investigating Cross-Cultural Pragmatics: An Introductory Overview'. In *Cross-Cultural Pragmatics: Requests and Apologies*, edited by Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper. Vol. Advances in discourse processes. Norwood, N.J.: Ablex.

Brown, Penelope, and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics 4. Cambridge: University Press.

Bybee, Joan, ed. 2010. 'Conventionalization and the Local vs. the General: Modern English Can'. In *Language, Usage and Cognition*, 151–64. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511750526.009.

Cai, Zhenguang G., David A. Haslett, Xufeng Duan, Shuqi Wang, and Martin J. Pickering. 2023. 'Does ChatGPT Resemble Humans in Language Use?' arXiv. https://doi.org/10.48550/arXiv.2303.08014.

Chen, Xi. 2022. 'The Pragmatic Resistance of Chinese Learners of Korean'. *Foreign Language Annals* 55 (4): 1128–51. https://doi.org/10.1111/flan.12656.

Chen, Xi, and Lucien Brown. 2022. 'Second Language Knowledge of Pragmatic Meanings: What Do Learners of Korean Know about the Korean Pronouns Ce and Na?' *Journal of Pragmatics* 202 (December): 7–22. https://doi.org/10.1016/j.pragma.2022.10.010.

Chen, Xi, and Wei Ren. 2023. 'Functions, Sociocultural Explanations and Conversational Influence of Discourse Markers: Focus on Zenme Shuo Ne in L2 Chinese'. *International Review of Applied Linguistics in Language Teaching*, May. https://doi.org/10.1515/iral-2022-0230.

Chen, Xi, and Jiayi Wang. 2021. 'First Order and Second Order Indirectness in Korean and Chinese'. *Journal of Pragmatics* 178 (June): 315–28. https://doi.org/10.1016/j.pragma.2021.03.022.

Culpeper, Jonathan, and Vittorio Tantucci. 2021. 'The Principle of (Im)Politeness Reciprocity'. *Journal of Pragmatics* 175 (April): 146–64. https://doi.org/10.1016/j.pragma.2021.01.008.

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.

Cheng, Tsui-Ping. 2016. 'Authentic L2 Interactions as Material for a Pragmatic Awareness-Raising Activity'. *Language Awareness* 25 (3): 159–78. https://doi.org/10.1080/09658416.2016.1154568.

Cohen, Andrew D. 1996. 'Developing the Ability to Perform Speech Acts'. *Studies in Second Language Acquisition* 18 (02): 253–67. https://doi.org/10.1017/S027226310001490X.

Cook, Vivian. 1999. 'Going Beyond the Native Speaker in Language Teaching'. *TESOL Quarterly* 33 (2): 185–209. https://doi.org/10.2307/3587717.

Cunningham, D. Joseph. 2017. 'Second Language Pragmatic Appropriateness in Telecollaboration: The Influence of Discourse Management and Grammaticality'. *System*, Special issue on Telecollaboration, 64 (February): 46–57. https://doi.org/10.1016/j.system.2016.12.006.

Delin, Judy, Anthony Hartley, and Donia Scott. 1996. 'Towards a Contrastive Pragmatics: Syntactic Choice in English and French Instructions'. *Language Sciences* 18 (3): 897–931. https://doi.org/10.1016/S0388-0001(96)00054-X.

Economidou-Kogetsidis, Maria. 2013. 'Strategies, Modification and Perspective in Native Speakers' Requests: A Comparison of WDCT and Naturally Occurring Requests'. *Journal of Pragmatics* 53 (July): 21–38. https://doi.org/10.1016/j.pragma.2013.03.014.

———. 2016. 'Variation in Evaluations of the (Im)Politeness of Emails from L2 Learners and Perceptions of the Personality of Their Senders'. *Journal of Pragmatics* 106 (December): 1–19. https://doi.org/10.1016/j.pragma.2016.10.001.

Edmonds, Amanda. 2014. 'Conventional expressions: Investigating Pragmatics and Processing'. *Studies in Second Language Acquisition* 36 (1): 69–99.

Eelen, Gino. 2001. *A Critique of Politeness Theories*. St. Jerome Pub.

Fatemeh, Moafian, Yazdi Naji, and Sarani Abdulah. 2021. 'A Gendered Study of Refusal of Request Speech Act in the Three Languages of Persian, English and Balouchi: A within Language Study'. *International Review of Applied Linguistics in Language Teaching* 59 (1): 55–85. https://doi.org/10.1515/iral-2017-0084.

Fraser, Bruce. 2010. "Pragmatic Competence: The Case of Hedging." In *New Approaches to Hedging*, edited by Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider, 15–34. Emerald.

Fukushima, Saeko. 1996. 'Request Strategies in British English and Japanese'. *Language Sciences* 18 (3–4): 671–88. https://doi.org/10.1016/S0388-0001(96)00041-1.

Giray, Louie. 2023. 'Prompt Engineering with ChatGPT: A Guide for Academic Writers'. *Annals of Biomedical Engineering* 51 (12): 2629–33. https://doi.org/10.1007/s10439-023-03272-4.

Gumperz, John J. 1982. *Discourse Strategies*. Cambridge University Press.

Herbold, Steffen, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. 'AI, Write an Essay for Me: A Large-Scale Comparison of Human-Written versus ChatGPT-Generated Essays'. arXiv. https://doi.org/10.48550/arXiv.2304.14276.

Hoshi, Saori. 2022. 'Effects of Classroom Instruction on the Development of L2 Interactional Resource for Joint Stance Taking: Use of Japanese Interactional Particle Yo in Spontaneous Peer Conversation'. *Applied Linguistics* 43 (4): 698–724. https://doi.org/10.1093/applin/amab074.

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.

Hymes, Dell. 1966. 'Two Types Of Linguistic Relativity'. In William Bright (ed.). *Sociolinguistics*, 114–67. The Hague: Mouton. Paris

Ishihara, Noriko. 2008. 'Transforming Community Norms: Potential of L2 Speakers' Pragmatic Resistance', 11.

Ishihara, Noriko, and Andrew D. Cohen. 2010. *Teaching and Learning Pragmatics: Where Language and Culture Meet*. Harlow: Longman.

Ji, Hyangeun, Insook Han, and Yujung Ko. 2023. 'A Systematic Review of Conversational AI in Language Education: Focusing on the Collaboration with Human Teachers'. *Journal of Research on Technology in Education* 55 (1): 48–63. https://doi.org/10.1080/15391523.2022.2142873.

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.

Kinginger, Celeste, and Kathleen Farrell. 2004. 'Assessing Development of Meta-Pragmatic Awareness in Study Abroad'. *Frontiers: The Interdisciplinary Journal of Study Abroad* 10 (1): 19–42. https://doi.org/10.36366/frontiers.v10i1.131.

Kwon, Jihyun. 2004. 'Expressing Refusals in Korean and in American English'. *Multilingua* 23 (4): 339–64. https://doi.org/10.1515/mult.2004.23.4.339.

Laughlin, Veronika Timpe, Jennifer Wain, and Jonathan Schmidgall. 2015. 'Defining and Operationalizing the Construct of Pragmatic Competence: Review and Recommendations'. *ETS Research Report Series* 2015 (1): 1–43. https://doi.org/10.1002/ets2.12053.

Lazare,M.D, Aaron. 2005. *On Apology*. Oxford, New York: Oxford University Press.

Leech, Geoffrey. 1983. *Principles of pragmatics*. London, New York: Longman Group Ltd.

Li, Yuan ke, Shiwan Lin, Yarou Liu, and Xiaofei Lu. 2023. 'The Predictive Powers of Fine-Grained Syntactic Complexity Indices for Letter Writing Proficiency and Their Relationship to Pragmatic Appropriateness'. *Assessing Writing* 56 (April): 100707. https://doi.org/10.1016/j.asw.2023.100707.

Liao, Wenxiong, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, et al. 2023. 'Differentiate ChatGPT-Generated and Human-Written Medical Texts'. arXiv. https://doi.org/10.48550/arXiv.2304.11567.

Loconte, Riccardo, Graziella Orrù, Mirco Tribastone, Pietro Pietrini, and Giuseppe Sartori. 2023. 'Challenging ChatGPT ' *Intelligence*' with Human Tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model'. SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4377371.

Ma, Yongqiang, Jiawei Liu, Fan Yi, Qikai Cheng, Yong Huang, Wei Lu, and Xiaozhong Liu. 2023. 'AI vs. Human -- Differentiation Analysis of Scientific Content Generation'. arXiv. https://doi.org/10.48550/arXiv.2301.10416.

Mills, Sara, and Karen Grainger. 2016. *Directness and Indirectness Across Cultures*. Springer.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Analytical Methods for Social Research. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107587991.

Nelson, Gayle L., Mahmoud Al Batal, and Waguida El Bakary. 2002. 'Directness vs. Indirectness: Egyptian Arabic and US English Communication Style'. *International Journal of Intercultural Relations* 26 (1): 39–57. https://doi.org/10.1016/S0147-1767(01)00037-2.

Purpura, James E. 2004. *Assessing Grammar*. Cambridge Language Assessment. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511733086.

Qiu, Zhuang, Xufeng Duan, and Zhenguang Garry Cai. 2023. 'Pragmatic Implicature Processing in ChatGPT'. PsyArXiv. https://doi.org/10.31234/osf.io/qtbh9.

Roever, Carsten. 2011. 'Testing of Second Language Pragmatics: Past and Future'. *Language Testing* 28 (4): 463–81. https://doi.org/10.1177/0265532210394633.

Rosen, Larry D., Jennifer Chang, Lynne Erwin, L. Mark Carrier, and Nancy A. Cheever. 2010. 'The Relationship Between "Textisms" and Formal and Informal Writing Among Young Adults'. *Communication Research* 37 (3): 420–40. https://doi.org/10.1177/0093650210362465.

Rosenbaum, Paul.R. 2002. *Observational Studies. Springer Series in Statistics*. Springer, New York, NY.

Shao, Kaiqi, Reinhard Pekrun, and Laura J. Nicholson. 2019. 'Emotions in Classroom Language Learning: What Can We Learn from Achievement Emotion Research?' *System*, Special Issue: New directions for individual differences research in language learning, 86 (November): 102121. https://doi.org/10.1016/j.system.2019.102121.

Shishavan, Homa Babai, and Farzad Sharifian. 2013. 'Refusal Strategies in L1 and L2: A Study of Persian-Speaking Learners of English'. *Multilingua* 32 (6): 801–36. https://doi.org/10.1515/multi-2013-0038.

Su, Hang. 2017. 'Local Grammars of Speech Acts: An Exploratory Study'. *Journal of Pragmatics* 111 (April): 72–83. https://doi.org/10.1016/j.pragma.2017.02.008.

Su, Hang, and Yixin Fu. 2023. 'Local Grammar Approaches to Speech Acts in Chinese: A Case Study of Exemplification'. *Journal of Pragmatics* 212 (July): 44–57. https://doi.org/10.1016/j.pragma.2023.05.004.

Su, Yunwen, and Wei Ren. 2017. 'Developing L2 Pragmatic Competence in Mandarin Chinese: Sequential Realization of Requests'. *Foreign Language Annals* 50 (2): 433–57. https://doi.org/10.1111/flan.12263.

Takahashi, Tomoko, and Leslie M. Beebe. 1987. 'The Development of Pragmatic Competence by Japanese Learners of English'. JALT Journal 8.2. 131-155.

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.

Tantucci, V., Wang, A., & Culpeper, J. (2022). Reciprocity and epistemicity: On the (proto)social and cross-cultural 'value' of information transmission. *Journal of Pragmatics*, *194*, 54–70. https://doi.org/10.1016/j.pragma.2022.04.012

Thomas, Jenny. 1983. 'Cross-Cultural Pragmatic Failure'. *Applied Linguistics* 4 (2): 91–112. https://doi.org/10.1093/applin/4.2.91.

van Compernolle, Rémi A. 2014. 'Sociocultural Theory and L2 Instructional Pragmatics'. In *Sociocultural Theory and L2 Instructional Pragmatics*. Multilingual Matters. https://doi.org/10.21832/9781783091409.

Van Dijk, Teun A. 1979. 'Pragmatic Connectives'. *Journal of Pragmatics* 3 (5): 447–56. https://doi.org/10.1016/0378-2166(79)90019-5.

Van Ek, Jan Ate. 1986. *Objectives for Foreign Language Learning*. Council of Europe.

Yu, Kyong-Ae. 2011. 'Culture-Specific Concepts of Politeness: Indirectness and Politeness in English, Hebrew and Korean Requests'. *Intercultural Pragmatics* 8 (3): 385–409. https://doi.org/10.1515/iprg.2011.018.