A Quantitative Verification of Politeness Theory Using Large Language Models from Japanese Dialogue Data

Tetsuro Takahashi[1] Mayumi Usami[2]
[1]Kagoshima University [2]Tokyo University of Foreign Studies
takahashi@ibe.kagoshima-u.ac.jp usamima@tufs.ac.jp

Abstract

The primary objective of this study is to quantitatively validate Brown and Levinson's politeness theory, which has garnered significant attention and extensive research within the realm of social sciences. Using a large language model (LLM), we estimated the politeness level of utterances and the rank of imposition on the listener within a Japanese dialogue corpus. By comparing face-threatening acts with politeness levels, we found a significant correlation in 98 out of 120 dialogues analyzed in our experiment. These results confirm the validity of the politeness theory as observed in these dialogues.

1 Introduction

Large-scale corpora and LLMs have advanced dialogue systems, enabling natural conversations with humans[1, 2, 3]. Evaluations typically focus on task understanding, logical consistency, and coherence but overlook aspects studied in linguistics and social sciences, such as language variations, politeness strategies, and interpersonal dynamics. Interpersonal dynamics in dialogue research remain insufficiently integrated in system development. As a result, while dialogue systems manage tasks and casual conversations, they still require users to adapt. Previously, integrating dialogue research into systems was challenging, but LLMs now offer the potential to bridge this gap, incorporating established theories into practical applications.

As part of this initiative, our study focuses on politeness theory[4] and addresses the following:

- Estimate values related to politeness theory, along with the politeness level of utterances, using LLMs on a dialogue corpus.
- Quantitatively verify politeness theory by comparing face-threatening acts with politeness levels in a dialogue corpus.

2 Politeness Theory

Politeness theory is a linguistic framework that systematizes strategies for establishing and maintaining smooth human relationships. Brown and Levinson (henceforth B&L)[4] introduced definitions that have had a significant impact on subsequent research.

B&L proposed the concept of *face* in relation to politeness in communication:

Positive Face The desire to be understood, liked, and appreciated by others.

Negative Face The desire to be free from interference and intrusion by others.

They argued that communication inevitably involves actions that threaten the listener's face, and that politeness serves as a linguistic strategy to mitigate such face-threatening acts. The degree of face threat (W_x) is estimated using the following equation:

$$W_x = D(S, H) + P(H, S) + R_x \tag{1}$$

W_x: Degree of face threat, indicating how much an act (x) threatens the listener's face.

D: Social distance between the speaker (S) and the hearer (H).

P: Power of the hearer (H) over the speaker (S).

R_x: rank of imposition: cultural weight assigned to an act (x) based on the burden it places on the listener.

One implication of this equation is that *higher face threat (W_x) leads to more polite expressions.* This study quantitatively examines this phenomenon using a dialogue corpus.

3 Estimating Face Threat and Politeness with LLMs

3.1 Data Used

We used the BTSJ1000 Japanese natural conversation corpus[5]. This corpus contains 514 dialogues between speakers with varying relationships, including strangers, friends, and teacher-student, making it a valuable resource for linguistic research. From this corpus, we selected 120 dialogues from folders #1, #2, #3, #4, #5, #17, and #18 for our analysis.

3.2 Estimation Using LLMs

In this study, we estimated the degree of face-threatening acts (W_x) and politeness (Po) in utterances within a dialogue corpus.

First, the value of social distance (D) between speakers and the power (P) of the speaker were defined by hand based on the information about speakers described in the corpus as follows.

D		P	
known	3	unknown	3
unknown	5	friend	3
		Speaker is older	4
		Speaker is younger	2
		Speaker is teacher	5
		Speaker is student	1

The values were defined as a five-scale score ranging from 1 to 5. A value of 3 was assigned to social distance (D) when participants knew each other, and 5 when they met for the first time. For power (P), a value of 1 was assigned to the student and 5 to the teacher, reflecting the significant power imbalance in their relationship. In cases where the speaker is younger or older than the listener, values of 2 and 4 were assigned, respectively, to account for power differences, particularly in the context of Japanese culture. In other cases, such as relationships between strangers or friends, a value of 3 was assigned to power.

We used GPT-4o-mini model provided by OpenAI[6] as the LLM to estimate the rank of imposition (R_x) and the degree of politeness (P_o) with the prompts described in Figure 1 and Figure 2 in Appendix. For all 25,844 utterances across 120 dialogues, the face-threatening degree (W_x) was calculated as the sum of social distance (D), power (P) and rank of imposition (R_x). An example of the estimation results is shown in Table 1. This dialogue is a request interaction where Speaker 1 makes a request to Speaker 2. Regarding R_x, backchannel responses such as "Uh-huh" or "Eh" were estimated with a low value of 1, whereas utterances related to making a request were estimated with higher values. Since backchannel responses impose little burden on the listener while requests impose a greater burden, these estimations appear to be reasonable.

Next, looking at the estimated results for politeness (P_o), neutral expressions such as "Next Monday" or "At 9 AM" received mid-range values of 3 on a scale from 1 to 5. Meanwhile, casual expressions like "Uh-huh" or "Eh" received low values, as they are highly informal. Conversely, a highly polite utterance, such as "Would you be willing to participate in an experiment related to language research?"–which is in formal and respectful language–was assigned a high score of 5. These results suggest that the estimation of politeness (P_o) is also reasonable.

Table 1: Estimation of Face-Threatening Degree and Politeness by LLM

Speaker	D	P	R_x	W_x	Po	Utterance
1	3	4	2	9	3	I called because I have a request.
2	3	2	1	6	1	Uh-huh.
1	3	4	1	8	3	Next Monday.
2	3	2	1	6	1	Uh-huh.
1	3	4	1	8	3	At 9 AM.
2	3	2	1	6	1	Uh-huh.
1	3	4	3	10	3	You need to go to NINJAL Lab.
2	3	2	1	6	1	Uh-huh.
1	3	4	4	11	3	On my behalf.
2	3	2	1	6	1	Uh-huh.
1	3	4	4	11	5	Would you be willing to participate in an experiment related to language research?
2	3	2	1	6	1	Eh, huh, heh, eh.
1	3	4	3	10	3	Moreover, with a Korean person.

Table 2: Statistical Values for Dialogue Categories Used in BTSJ
("#Dialog*" represents the number of dialogues where W_x and Po are significantly correlated)

Dialogue Category	#Dialog	#Dialog*	Avg. W_x	Avg. Po	Correl
Casual Between Friends (Male-Male)	10	9	1.63	2.31	0.56
Casual Between Friends (Female-Female)	21	19	1.66	2.29	0.70
First-time Meeting (Female-Female)	11	11	1.59	3.38	0.42
Thesis Guidance (Teacher-Student)	10	9	1.99	3.32	0.45
Refusal (To older; Female-Female)	13	13	1.56	2.84	0.62
Refusal (To Peer; Female-Female)	13	10	1.59	2.73	0.57
Refusal (To Junior; Female-Female)	13	6	1.60	3.19	0.17
Request Between Friends (Male-Male)	10	7	1.49	2.66	0.58
Request Between Friends (Female-Female)	10	9	1.44	2.61	0.53
Debate Between Friends (Mixed-Gender)	5	5	1.78	2.66	0.75
First-time Meeting Debate (Female-Female)	4	4	2.33	3.49	0.12

Based on the above results, we conclude that the LLM is able to estimate both face-threatening degree (W_x) and politeness (P_o) with reasonable accuracy.

4 Analysis of Estimation Results

We estimated the degree of face-threatening acts (W_x) and politeness (P_o) for all utterances in the 120 dialogues included in the target corpus and presented statistical values for each category. The results are shown in Table 2. Among the 120 dialogues, 98 dialogues (81.7%) showed a significant correlation between W_x and Po, confirming that the formula for estimating the degree of face-threatening acts holds true for many real conversations. Because the data include information about the gender of speakers, we examined gender differences; however, no significant differences were found.

5 Conclusion

In this study, we estimated the degree of politeness and the rank of imposition in utterances using an LLM, based on the utterance content. We found that the face-threatening degree (W_x) and the politeness level (P_o) were significantly correlated. While each utterance was analyzed independently, they are naturally interrelated. A promising direction for future research is to extend this work to Discourse Politeness Theory[7].

References

[1] Vojtěch Hudeček and Ondřej Dušek. Are LLMs all you need for task-oriented dialogue? **arXiv preprint arXiv:2304.06556**, 2023.

[2] Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. PlatoLM: Teaching LLMs in Multi-Round Dialogue via a User Simulator. In **Proc. The Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7841–7863, 2024.

[3] Ryuichiro Higashinaka, Tetsuro Takahashi, Michimasa Inaba, Zhiyang Qi, Yuta Sasaki, Kotrao Funakoshi, Shoji Moriya, Shiki Sato, Takashi Minato, Kurima Sakai, Tomo Funayama, Masato Komuro, Hiroyuki Nishikawa, Ryosaku Makino, Hirofumi Kikuchi, and Mayumi Usami. Dialogue system live competition goes multimodal: Analyzing the effects of multimodal information in situated dialogue systems. In **Proc. The 14th International Workshop on Spoken Dialogue Systems Technology**, 2024.

[4] Penelope Brown and Stephen C Levinson. **Politeness: Some universals in language usage**. No. 4. Cambridge university press, 1987.

[5] Mayumi (Ed.) USAMI. Building of a japanese 1000 person natural conversation corpus for pragmatic analyses and its multilateral studies, and ninjal institute-based projects: Multiple approaches to analyzing the communication of japanese language learners., 2023.

[6] OpenAI. GPT-4o technical report. https://openai.com/index/gpt-4o, 2024. Accessed: 2024-12-29.

[7] Mayumi Usami. Discourse politeness theory and cross-cultural pragmatics. In **Readings in second language pedagogy and second language acquisition: In Japanese context**, pp. 19–41. John Benjamins Publishing Company, 2008.

A Prompt

In politeness theory, the weight of a face-threatening act (Wx) is formulated as follows:
Wx = D(S, H) + P(H, S) + Rx
Here, D, P, and Rx are defined as follows:
D (Social Distance): Based on the closeness or relationship between the speaker and the hearer.
P (Power): The influence or authority the hearer holds over the speaker.
Rx (Rank of Imposition): The degree of burden or impact the utterance imposes on the hearer. Ranges from 1 (little to no burden) to 5 (high burden).
For each of the following utterances, please estimate the value of Rx.
No explanation is needed; output only the Rx values.

Figure 1: Prompt for Rank of Imposition

For each of the following utterances, rate the level of politeness on a 5-point scale from 1 (not polite) to 5 (very polite), and output the ID and politeness score separated by a comma.
—
ID \t Speaker \t Utterance

Figure 2: Prompt for Politeness