The Effect of Response Suggestion on Dialogue Flow: Analysis Based on Dialogue Act and Initiative

Minami Inoue (inoue@ailab.ics.keio.ac.jp)

Tomoyuki Maekawa (maekawa@ailab.ics.keio.ac.jp)

Ryoichi Shibata (shibata@ailab.ics.keio.ac.jp)

Michita Imai (michita@ailab.ics.keio.ac.jp)
Keio University, Yokohama, Japan

Abstract

Technology to predict responses is a key element in human-to-human messaging that has increasingly been utilized to enable AI-mediated communication. When response suggestions from AI are incorporated into messaging between humans, it will have an impact on the flow and content of the dialogue. In particular, we cannot ignore the effect of the response suggestions generated by a large language model like GPT-3 that include information appropriate to a dialogue context beyond subsidiary responses generated by the previous language. In this paper, we investigated the effect of AI response suggestion sentences used for chat messaging on dialogue flow from two aspects: dialogue act and initiative. Usage rates of response suggestions for different dialogue acts were measured with BLEU scores, and we found that the response suggestions contributed to the establishment of a proper dialogue flow, such as an answer in response to a question. The results of our case study indicated that users who take the initiative in the dialogue tend to utilize response suggestions less frequently. We also found that some written responses were based on the suggested sentence structure but conveyed different messages.

Keywords: AI-mediated communication; dialogue act; dialogue initiative; response suggestion

Introduction

We are beginning to see a future where large-scale natural language processing models can replace human descriptions. For example, the emergence of ChatGPT (OpenAI, 2022), which can generate human-like sentences, has had a significant impact by attracting many users and bringing AI writing to the general public. Such AI may change the way people write.

In this study, we focus on AI-generated response suggestions. AI-generated response suggestions are commonly used in text-based communication. For example, Google's Smart Reply (Kannan et al., 2016), which provides AI-generated email reply suggestions, accounts for approximately 10% of all mobile replies. The response suggestion is one of the key elements of AI-Mediated Communication (AI-MC), which refers to "interpersonal communication in which an intelligent agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication goals" (Hancock, Naaman, & Levy, 2020).

As co-writing with AI is spreading, it is vital to examine the extent to which people rely on AI suggestions in conversation. Interestingly, several studies on AI co-writing have found that sentences suggested by AI may provide people with ideas (Bhat, Agashe, & Joshi, 2021; Osone, Lu, & Ochiai, 2021; Gero, Liu, & Chilton, 2022) and that the suggestions tend to make sentences contain more predictive words (Arnold, Chauncey, & Gajos, 2020). If AI language is incorporated into a dialogue, the goal of the dialogue can be changed. The goal may be the conclusion of a debate or getting to know the partner.

In the field of AI-MC, the effect of response suggestion has been studied from various perspectives. Hohenstein and Jung (2018) compared AI-supported and standard chat applications. Robertson, Olteanu, Diaz, Shokouhi, and Bailey (2021) detected inappropriate content for e-mail response suggestions. Buschek, Zürn, and Eiband (2021) investigated the effect of multiple input suggestions and modeled user input behavior. Hohenstein and Jung (2020) discussed trust and responsibility in AI-MC.

Mieczkowski, Hancock, Naaman, Jung, and Hohenstein (2021) provided piecemeal evidence that AI suggestions are mixed with human messages. They investigated the usage of smart replies in detail. Specifically, they investigated whether the positive bias contained in smart replies affected the messages sent, as well as how users incorporated AI suggestions into the response to the message from the partner during a conversation. They found that the messages were not significantly affected by the positive bias and that the users united the suggested words with their own words to make a sentence. This means that users created Human-AI composite messages that were a mixture of human and AI languages.

However, the effect of response suggestions on dialogue flow has not been fully investigated in terms of either dialogue acts or user factors. Because response suggestions are contextualized in the dialogue, they would tell how to proceed with the dialogue. As such, they may influence the dialogue acts of user responses. The effect of the user factors in the dialogue flow is related to the user's initiative. We speculate that the use of response suggestions in a dialogue is influenced by the intensity of the user's motivation toward a dialogue goal. In other words, a user who is willing to engage in a dialogue would not use response suggestions.

In this paper, we investigated how AI-generated response suggestions changed the dialogue flow in human-to-human messaging from two aspects: dialogue acts and user initiative. In Experiment 1, we compared AI language with human language and found that AI language was chosen as a response suggestion in approximately 1/3 of the cases. In Experiment

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

2, we compared the extent of editing AI suggestions for different dialogue acts, and found that AI suggestions could be used to establish a dialogue flow. We also conducted a case study as Experiment 3 and found that user initiative in a dialogue may affect the frequency of using response suggestions. In addition, we observed cases where suggested responses determined the sentence structure regardless of user initiative.

Experiment 1: Comparison of AI Language and Human Language as Response Suggestions in Chat

The purpose of this experiment is to clarify the following research question:

- RQ1: To what extent do people use predictive sentences generated by current large-scale language processing models as response suggestions in chat?

In this experiment, we compared human-generated sentences and AI-generated predictive sentences. Specifically, we showed a participant a dialogue in chat corpus after instructing her/him to read as one of the interlocutors of the conversation. Then, we asked the participant to answer whether s/he would refer to the response in the corpus or to the sentence predicted by AI for writing her/his own response.

Participants

Fifty participants were recruited through a crowdsourcing website. Ages ranged from 22 to 64 ($M = 38.9, SD = 9.08$). In terms of gender, 21 participants identified as female and 29 as male. The participants were paid JPY 150 after the experiment.

Material

In the experiment, participants read the first eight messages (four turns) of each session in the chat corpus and selected a sentence to use as a reference for creating the next response. We used the Japanese version of persona chat corpus[1] published by NTT, which was created by giving two people persona settings and instructing them to chat in an attempt to get to know each other (Sugiyama et al., 2022). We selected 500 dialogues from the persona chat corpus.

Experimental Conditions

We prepared two different sources of the sentences presented as a response suggestion as experimental conditions.

- Corpus condition: Ninth response of the dialogue session in the chat corpus—i.e., The next response following the four turns of the dialogue. We employed the first sentence of the response if the response consisted of multiple sentences to make the participant's decision easier.

- AI condition: The response prediction of the dialogue corpus generated by GPT-3 (Brown et al., 2020). GPT-3 receives input sentences as a prompt and then generates the next sentence. We use OpenAI API[2] to generate the AI condition sentences by GPT-3. Four turns of dialogue history were input to OpenAI. Only the first sentence of the output was used if there were multiple generated sentences.

Procedure

The experiment was conducted through a questionnaire form on the Internet. We divided 500 dialogue corpus into fifty forms to give each participant ten individual dialogue sessions. Participants first read the instruction when they accessed the form. The instruction presented the following situations: "Suppose you are talking with one of your friends via a chat app. This chat app has an assistive function that provides two response suggestions as candidates for what you will say next. You will be able to edit the response after choosing a response suggestion."

After reading this instruction, participants checked the dialogue history of each of the ten dialogue sessions and then selected one response from two options generated by Corpus or AI. In addition, they were blind to the conditions of the response generation. They then answered the following two open questions:

- Criteria: What criteria did you use in deciding which response suggestion to use?

- Demand: What kind of assistance do you think would be useful when sentences are displayed as response suggestions in chat, as in the case of our presented situation?

Prediction of Results

We expected the AI condition to be selected equally as often as the Corpus condition. This is because a variety of sentences can be regarded as appropriate responses in a dialogue, and any of them can be chosen according to the speaker's background and preferences.

Results and Discussion

Out of 500 cases, AI-generated responses were chosen in 163 cases (32.6%). Although this is not quite as high as we had predicted, it is still a substantial frequency. This result suggests that the AI-generated responses are to some extent contextual and natural, and can be used adequately as response suggestions.

Some of the answers to the open questions were consistent with previous findings on AI-MC. Some participants answered that they preferred positive and polite expressions, which is consistent with the finding that messages composed with smart replies include more positive sentiment than messages composed without smart replies (Mieczkowski et al., 2021).

As for the demand, one participant said "If the assist is displayed as a complete sentence, even if it can be edited later, it would be disrespectful to the partner because it is not my original words." While previous studies have pointed out that

[1] https://github.com/nttcslab/japanese-dialog-transformers

[2] https://openai.com/api/

word prediction and phrase prediction tend to play different roles in writing (Arnold, Gajos, & Kalai, 2016), this comment suggests that the roles of phrase prediction and sentence prediction may also differ from each other.

Experiment 2: Comparison of Usage of Response Suggestions by AI for Different Dialogue Acts

The purpose of this experiment is to clarify the following research questions:

- RQ2: Does the type of AI response suggestion in a dialogue affect the editing behaviors of humans when incorporating them into their own responses?
- RQ3: Do humans alter the type of AI response suggestions when incorporating them into their own responses?

In this experiment, we analyzed how participants edited responses suggested by AI on the basis of dialogue acts. The participants created their own responses after reading the chat history of the corpus and response suggestions generated by AI. Since we wanted to focus on dialogue acts rather than initiative, the existing corpus was used so that the participants had little initiative in the dialogue.

Participants

Fifty participants were recruited through a crowdsourcing website. Ages ranged from 29 to 63 ($M = 42.5, SD = 8.06$). In terms of gender, 20 participants identified as female and 30 as male. Participants were paid JPY 350 after the experiment.

Material

In this experiment, participants read the first eight messages (four turns) of each session in the dialogue corpus and then created their own responses by editing the suggested responses. Suggested responses were generated in the same way as for the AI condition in Experiment 1. The chat corpus was also the same as in Experiment 1.

Experimental Conditions

Experiment 2 focused on the dialogue act, which is the function of each utterance in a dialogue (Popescu-Belis, 2005). Specifically, we compared the extent to which the suggested responses were edited for different dialogue acts.

An annotator classified suggested responses into five categories: *question*, *answer*, *inform*, *directive*, and *commissive*. This classification was based on ISO standard (Bunt et al., 2010). Table 1 lists the classification criteria. Ten responses per dialogue act were used in the experiment, for a total of 50 responses.

Table 1: Classification criteria of dialogue acts.

Dialogue act	Definition
Question	Statements requesting information from the partner.
Answer	Statements in response to a question that provides new information to the partner.
Inform	Statements that provide information that the partner does not know.
Directive	Statements that have an impact on the future behavior of the partner.
Commissive	Statements that are a declaration about the speaker's future actions.

Procedure

The experiment was conducted through a questionnaire form on the Internet. Participants were assigned to either of five forms (ten participants for each) consisting of different dialogue examples. Each form covered two sessions for each of five dialogue acts of suggested responses, namely ten sessions in total. The participants accessed the questionnaire and read the description of the situation settings, which was basically the same as the one in Experiment 1 except that only one suggested response was shown. In each session, the participants read the dialogue history and the suggested response, and then they created their own responses. The answer field was pre-filled with the suggested response so that we could measure the degree of editing. The participants then answered how useful they found the suggested response on a 10-point scale and categorized their own responses into one of the five dialogue acts. After ten sessions, they answered open questions about the use of response suggestions. All data collected were valid, namely, ten edit data were collected for each of the 50 suggested responses.

Prediction of Results

We expected that the suggested responses would be edited differently depending on the dialogue act. For example, dialogue acts that affect the partner, such as *question* and *directive*, would be edited more significantly so as to better express the participants' own opinion in reference to the suggested response, while *answer* and *commissive* would not be edited so much because the contents of the responses are restricted by what the partner says.

We also expected that the dialogue acts would be changed less frequently when responding to a partner's request, such as *answer* or *commissive*, whereas other dialogue acts would be changed more frequently.

Results and Discussion

For each of the 100 responses (ten responses per session), we calculated BLEU scores for the responses generated by the participants (edited response) and the suggested response. The BLEU score measures the extent to which the edited response consisted of the same words and phrases as the suggested response on the n-gram basis. The BLEU score close to 100 means that most of the words in the suggested response

Table 2: Average BLEU score for each dialogue act.

Dialogue act	BLEU score
Question	21.76
Answer	61.92
Inform	18.75
Directive	28.86
Commissive	27.72

remain without changes, while the BLEU score close to 0 means that the suggested response is drastically changed.

Table 2 shows the average BLEU score for each dialogue act. The results indicate that the BLEU score for *answer* is the highest. This may be because the participants did not have clear opinions to tell in response to the questions presented by the corpus. We speculate that users would utilize more of the response suggestions when they do not have their own motivation for writing. In addition, a question requires an answer in the response. Although *commissive* can be a response to *directive*, the restriction by a question is stronger in the dialogue flow. This may be the reason why the BLEU score for *answer* was higher than that for *commissive*.

The change in the dialogue act of the edited responses from the suggested responses is represented as a state transition matrix in Fig. 1. As hypothesized, we can see here that *answer* tended to be used without changing the dialogue act. Moreover, *commissive* was often changed to *answer*, which is a response to a question.

On the other hand, contrary to our hypothesis, *question* was not changed to other dialogue acts in more than half of the cases. This suggests that humans may be biased to ask questions when *question* is presented as a suggestion. One example from this experiment is as follows. The partner was talking about disliking cold weather in the context. Then a response suggestion saying "Is it going to be hot this summer, too?" was presented. Although the suggested response was inconsistent with the context, five out of ten participants wrote some kind of questions as their own responses. Some participants added another sentence after the suggestion, while others changed it to ask if the partner is good at hot weather, and others modified it to ask if it was cold that winter as well. This may be because there is a more variety of contents appropriate for *question* compared to *answer*.

Lastly, we found that the dialogue act was changed in 52% of all responses. This indicates that humans are likely to create diverse types of responses even when the responses are suggested by AI.

Experiment 3: Case Study

The purpose of this experiment is to observe real interactions that cannot be obtained in a crowdsourcing experiment, and gain insights into the following research questions:

- RQ4: Does the frequency with which humans utilize response suggestions vary depending on the individual initiative to the conversation?
- RQ5: Do response suggestions create a dialogue flow by themselves?
- RQ6: Is the similarity between a user's dialogue goal and the suggested response important for the utilization of response suggestion?

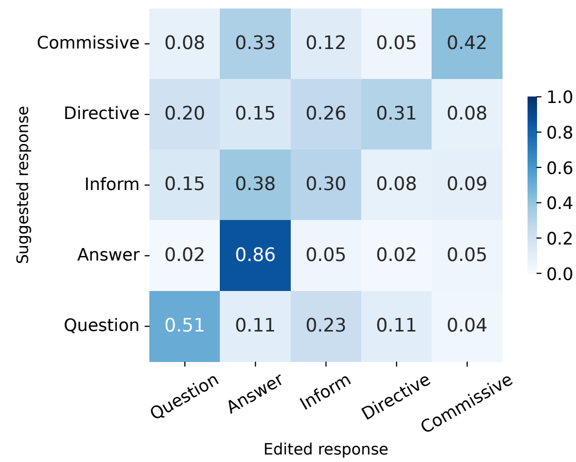

Figure 1: Transition of dialogue acts between the suggested responses and edited responses.

In this experiment, we asked users to try a chat application with AI-response suggestions and then examined the extent to which the dialogue flow changed with the suggestions. In addition, we interviewed the users to evaluate their initiative in the dialogue, namely whether they were willing to achieve certain dialogue goals.

Experimental Interface

As shown in Fig. 2, we implemented the interface of the chat application with response suggestions for this experiment. Up to three different suggestions are displayed above the input box. The user can click on one of the suggested responses and send the response as is or after editing it. Suggested responses were generated utilizing the same method as the other experiments.

Participants

Six male students participated in the experiment.

Procedure

The participants were divided into three pairs. Both participants in each pair met together in the laboratory. The experimenter instructed the participants on how to use the chat application. Then, the participants were asked to chat with each other using the application on individual laptops. After that, the participants were interviewed individually.

Figure 2: Screenshot of the chat application interface.

Each pair performed two chat sessions. The following two themes were set in the respective sessions:

- Free: The participants were allowed to talk about any topics.
- Trip: The participants were asked to make a fictional plan to go on a trip. They discussed the destination and activities.

The participants sent their responses alternately because the application did not accept successive inputs from one user. Videos of the application were recorded during the sessions.

The participants watched the recorded videos and answered the post-experiment interview. Specifically, they answered the following items:

- What they had thought about each response suggestion
- How they had created each of their actual responses
- How they had generally used the response suggestions

The answers to the interview were noted by the experimenter.

Results and Discussion

Table 3 lists the total number of responses, BLEU scores when the response suggestion was utilized, and suggestion usage rate for each session. A total of 81 responses were sent, of which 35 were created using the response suggestions. The average length of each response was 31.13 characters in Japanese. We describe the events related to each research question.

Individual Initiative to the Conversation As shown in Table 3, the suggestion usage rates vary widely among the participants. Participant F, who did not use the assistants at all, explained the reason that "I didn't rely on the suggestions because I wanted to talk by myself." This implies that if the users have a higher initiative in the conversation, they tend to utilize the response suggestions less frequently.

Table 3: Session information.

Case	Theme	Users	No. of turns	Suggestion usage rate	BLEU
1	Free	A & B	13	0.54	17.51
2	Trip	A & B	13	0.46	16.30
3	Free	C & D	15	0.53	48.00
4	Trip	C & D	17	0.76	44.17
5	Free	E & F	12	0.08	10.41
6	Trip	E & F	11	0.00	0.00

The Role of Suggested Response in Dialogue Flow In the interview, we asked the participants in which cases they had used response suggestions throughout the session (multiple answers were allowed):

- Self-message: The suggestion was what I wanted to say.
- Idea: The suggestion was good and I had not yet decided what I wanted to say.
- Change: I had something I wanted to say but changed my mind after seeing the response suggestion.

Four participants answered that they used the suggestions as their self-messages. Participant C stated that he used it to reduce the input effort. In particular, he often utilized a suggested response as a short response and then added his own sentence. In his case, the suggestion was used as a tool for input, and the dialogue flow was determined by himself.

Three participants answered that they used the suggestions as ideas. Table 4 shows an example that a participant used the suggestion because he could not come up with a response. Participant D commented "I sometimes changed my mind when I saw the suggestions, but I mainly looked for reasonable suggestions. So I did not always use the suggestions." This can be taken as either the suggestion being used as an idea or changing his mind. In this case, response suggestions presented new dialogue goals for users who do not have clear dialogue goals.

In addition, no participant clearly answered that the suggestions changed their mind. Interestingly, several participants said that they do not look at suggestions when they had decided what to say.

Incorporation of Suggestions into Own Responses Using the created responses and the interviews as a basis, we examine the impact of the similarity between participants' dialogue goals and the suggested responses on how they used the response suggestions.

We observed cases in which suggestions were used as a part of a sentence to express the message. Interestingly, even when the suggested responses were significantly edited, the

Table 4: Dialogue example in which a user created a response based on the suggested dialogue flow (originally in Japanese; translated into English for inclusion in this paper).

Context	D: Hello. Nice to meet you today!
Suggested Response	"Do you have any plans today?"
Actual Response	C: Do you have any plans after this experiment?

Table 5: Dialogue example in which a suggested response was significantly edited (originally in Japanese; translated into English for inclusion in this paper).

Context	B: It's so true, I always wonder why I'm doing this research. It's different from the past in that we can't just create a system and be done with it.
Suggested Response	"Well, I definitely have some thoughts."
Actual Response	A: Definitely. But we've got about a month to go, so let's keep our spirits up!

participants often used the suggestions because the suggestion matched their opinions. In the example shown in Table 5, the suggestion was adopted because it matched the participant's own opinion, but the actual response was very different from the suggestion. However, since the actual response used the word "definitely," which is also included in the suggestion, it is likely that the content of the suggested response influenced the word choice to express what he wanted to say.

We also observed a case in which the sentence structure of the suggested response was used even when the suggestion was different from what the user wanted to say. A typical example is shown in Table 6. Participant D told us "I used it because it would be true if I rewrote part of the sentence." Unlike the previous example, this case is an example that a part of the suggested response was used as a template after thinking of what to say.

General Discussion

Implication

In this section, we comprehensively discuss the impact of AI-MC on human writing.

In this paper, we obtained the following results from the experiments. In Experiment 2, we investigated the impact of the suggested response in terms of dialogue acts and found that suggestions of *answer* was used more frequently than that of the other dialogue acts. This implies that the response suggestions have an effect on establishing a dialogue flow, such as responding to questions posed. In the case study, the response suggestions had very little influence on the initiative a user already had. On the other hand, response suggestions had an impact on the responses being written in terms of word and sentence structure. There were also cases where response suggestions were utilized as ideas when the user did not have the initiative in the dialogue.

Table 6: Dialogue example in which a response suggestion was used as a sentence template (originally in Japanese; translated into English for inclusion in this paper).

Context	C: Teaching people programming sounds amazing! What exactly do you teach?
Suggested Response	"I mainly teach HTML, CSS, and JavaScript to learners."
Actual Response	D: I mainly teach Python to learners.

Our findings demonstrate that in AI-MC, AI assists the user in proceeding with the desired dialogue flow in the conversation. Response suggestions assist people in better verbalizing what they want to say by providing response formats, words, phrases, and ideas. If the user already has a dialogue goal, AI can assist the user in reaching that goal through the suggestions it provides. On the other hand, if a user does not have a clear dialogue goal, AI is used to present various potential dialogue flows aimed at keeping the conversation going. In this case, AI may set a new dialogue flow and thereby affect the goal of the dialogue. In both cases, however, AI is expected to assist the user while respecting the user's own initiative.

Limitations

There are several limitations to this study. First, the results were not analyzed with inferential tests. Second, the number of participants in Experiment 3 is only six and they are all males. Third, the classification criteria of dialogue acts may differ between the annotator and editors. Lastly, this study was conducted in Japanese, and the use of suggestions will almost certainly differ depending on the language.

Comparison with Previous Studies

This is the first study to apply dialogue acts and BLEU scores to investigate the degree of language blending in the field of AI-MC. One prior research examined human-AI composite messaging (Mieczkowski et al., 2021) but focused on whether the dialogue was edited, appended, or not. Our study presents a more detailed investigation of the editing rate. Moreover, as the response suggestions in this study were generated by GPT-3, we examined the effectiveness of suggestions that are more embedded in the human context compared to the existing co-writing studies that were based on GPT-2 (Buschek et al., 2021; Bhat et al., 2021; Gero et al., 2022).

Acknowledgments

This work was supported by JST CREST Grant number JP-MJCR19A1, Japan.

References

Arnold, K. C., Chauncey, K., & Gajos, K. Z. (2020). Predictive text encourages predictable writing. In *Proceedings of the 25th international conference on intelligent user interfaces* (p. 128–138). New York, NY, USA: Association for Computing Machinery.

Arnold, K. C., Gajos, K. Z., & Kalai, A. T. (2016). On suggesting phrases vs. predicting words for mobile text composition. In *Proceedings of the 29th annual symposium on user interface software and technology* (p. 603–608). New York, NY, USA: Association for Computing Machinery.

Bhat, A., Agashe, S., & Joshi, A. (2021, April). How do people interact with biased text prediction models while writing? In *Proceedings of the first workshop on bridging human–computer interaction and natural language processing* (pp. 116–121). Online: Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., ... Traum, D. (2010, May). Towards an ISO standard for dialogue act annotation. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Buschek, D., Zürn, M., & Eiband, M. (2021). The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery.

Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for science writing using language models. In *Designing interactive systems conference* (p. 1002–1019). New York, NY, USA: Association for Computing Machinery.

Hancock, J. T., Naaman, M., & Levy, K. (2020, 01). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89-100.

Hohenstein, J., & Jung, M. (2018). Ai-supported messaging: An investigation of human-human text conversation with ai support. In *Extended abstracts of the 2018 chi conference on human factors in computing systems* (p. 1–6). New York, NY, USA: Association for Computing Machinery.

Hohenstein, J., & Jung, M. (2020, may). Ai as a moral crumple zone: The effects of ai-mediated communication on attribution and trust. *Comput. Hum. Behav.*, *106*(C).

Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., ... Ramavajjala, V. (2016). Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (p. 955–964). New York, NY, USA: Association for Computing Machinery.

Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M., & Hohenstein, J. (2021). Ai-mediated communication: Language use and interpersonal effects in a referential communication task. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–14.

OpenAI. (2022). *Chatgpt: Optimizing language models for dialogue.* (https://openai.com/blog/chatgpt/)

Osone, H., Lu, J.-L., & Ochiai, Y. (2021). Buncho: Ai supported story co-creation via unsupervised multitask learning to increase writers' creativity in japanese. In *Extended abstracts of the 2021 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery.

Popescu-Belis, A. (2005). Dialogue acts: One or more dimensions. *ISSCO WorkingPaper*, *62*, 1–46.

Robertson, R. E., Olteanu, A., Diaz, F., Shokouhi, M., & Bailey, P. (2021). "i can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the 2021 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery.

Sugiyama, H., Mizukami, M., Arimoto, T., Narimatsu, H., Chiba, Y., Nakajima, H., & Meguro, T. (2022). Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 ieee spoken language technology workshop (slt)* (p. 4-1-17-TLP).