The Relationship Between Dialogue Acts and Idea Generation in Human–Human Collaborative Story Writing

Natsumi Ezure and Michimasa Inaba
The University of Electro-Communications
Chofu, Tokyo, Japan
e2430013@edu.cc.uec.ac.jp, m-inaba@uec.ac.jp

Abstract

With the rapid advancement of generative models, human–AI co-creation systems have been widely studied. However, over-reliance on AI-generated ideas may reduce original human ideas. To inform the design of AI systems that better elicit human creativity, we investigate human–human co-creative dialogue, focusing on dialogue acts (DAs) that promote idea generation. In this study, we collected data on collaborative story creation between two human workers assigned to asymmetric roles: a leader, who writes the story, and a supporter, who assists via chat. The dataset comprises 485 dialogues annotated with dialogue acts, presurvey and postsurvey evaluations. Logistic regression analysis results revealed that the leader's self-assessed idea quantity decreased as their perception of the supporter's idea quantity increased. Furthermore, correlation analysis showed that a higher frequency of accepting proposals and positive opinions from supporters was positively correlated with a higher number of idea proposals from the leader.

1 Introduction

The development of AI technologies, including Large Language Models (LLMs) trained on vast text data, has expanded possibilities for co-creation that combines human ideas and experience with AI. Human–AI co-creation systems have been widely studied; in these systems, users interact with AI to collaboratively work on creative tasks, often by issuing commands for partial modifications or the regeneration of AI-generated content (Oh et al., 2018; Davis et al., 2016; Huang et al., 2020; Louie et al., 2020; Kumaran et al., 2023). Furthermore, systems that interact with people through natural language have been investigated (Schmitt and Buschek, 2021; Yuan et al., 2022). The human–AI co-creation process has been shown to consist of three stages: Ideation, Illumination, and Implementation (Wan et al., 2024). LLMs are frequently employed in the initial Ideation phase. However, a notable issue in ideation tasks is that users tend to accept AI suggestions (Qin et al., 2025). Koivisto and Grassini (2024) found that the best human ideas still matched or exceeded ideas of chatbots. These findings indicate that eliciting ideas from humans is important for co-creation. In contrast, during human–human collaborative dialogue, a partner does not always provide an abundance of ideas. In fact, a situation where a partner offers fewer ideas might motivate an individual to generate more of their own, fostering self-driven ideation. From this perspective, identifying the characteristics of dialogue that promote idea generation in human–human interaction could offer new insights for designing co-creative AI that more effectively elicits human creativity. This study aims to identify the features of utterances that facilitate idea generation by analyzing dialogue data from a human–human co-creative process. Specifically, we address the following research questions:

- **RQ1:** In a co-creative dialogue, what is the relationship between dialogue acts and the quantity of ideas generated, as well as the evaluation of the partner?
- **RQ2:** In a co-creative dialogue, what is the relationship between one's own dialogue acts and the partner's dialogue acts?

In this study, we collected data on collaborative story creation between two human workers exchanging ideas. In our experimental setting, two workers were assigned asymmetric roles: a leader, who was responsible for writing the story on an interface, and a supporter, who could only assist through text chat. Logistic regression analysis results revealed that the leader's self-assessed idea quantity decreased as their perception of the supporter's idea quantity increased. Furthermore, correlation analysis revealed that the number of idea

proposals from leaders increased when supporters offered more frequent acceptances and positive opinions.

2 Related Work

2.1 Human–AI Collaborative Creation

Several studies have been conducted on human–AI co-creation systems using generative models in various domains, including drawing (Oh et al., 2018; Davis et al., 2016) and music (Huang et al., 2020; Louie et al., 2020).

Specifically, various studies have been conducted on collaborative text writing. BunCho (Osone et al., 2021) is a plot co-creation system for Japanese novelists that generates plots based on user-inputted themes and keywords. Dramatron (Mirowski et al., 2023) is a collaborative scriptwriting system built using Chinchilla (Hoffmann et al., 2022). Dramatron first generates a story summary and then progressively generates the title, characters, plot overview, location descriptions, and dialogue for each scene. Users can instruct regeneration or manual correction at any time. CritiCS (Bae and Kim, 2024) is a framework where, after a user inputs an initial draft, multiple LLM critics and one human leader incrementally refine drafts of the plan and story over multiple rounds.

CharacterChat (Schmitt and Buschek, 2021) and Wordcraft (Yuan et al., 2022) are human–AI collaborative writing systems that allow for dialogue through natural language. CharacterChat (Schmitt and Buschek, 2021) is designed for users to create fictional characters while chatting with the system. Wordcraft (Yuan et al., 2022) is a text editor where users can write short stories with a large language model (LLM). Wordcraft users can generate and modify narratives while interacting with LLM in natural language. Similar to CharacterChat and Wordcraft, our work focuses on building a story through dialogue. However, our study distinguishes itself by analyzing human-human collaborative interactions for AI system design.

2.2 Idea Generation

Wan et al. (2024) found that a three-stage iterative Human–AI co-creativity process emerges in collaborative prewriting, consisting of Ideation, Illumination, and Implementation. However, previous research for ideation tasks suggests that humans may overly adopt LLM-generated ideas. For example, Qin et al. (2025) investigated the impact of the timing of LLM assistance on ideation tasks and found that using LLMs from the beginning reduced the number of original ideas. This suggests that AI-generated ideas can ultimately constrain human creativity. In an exercise measuring creativity in divergent thinking, Koivisto and Grassini (2024) found that the best human ideas still matched or exceeded ideas of chatbots, highlighting the importance of eliciting human creativity.

Building on this background, our study focuses specifically on the quantity of ideas to analyze the relationship between collaborative dialogue and creativity.

2.3 Dialogue Acts

Dialogue acts (DAs) are semantic tags assigned to utterances in dialogue (e.g., "suggest" or "answer"). DAs are widely used in dialogue analysis. Although the analysis of DAs in co-creative dialogue data remains limited, for example, Katuka et al. (2022) analyzed the relationship between DAs and participant satisfaction in dialogues. This study, in contrast, aims to identify DAs that may facilitate idea generation by analyzing the quantitative relationship between DAs and the number of ideas generated.

LLMs are used for various tasks in text analysis, such as summarization and classification, and have also been applied to the annotation of DAs. However, it has been reported that the performance of LLMs is inferior to that of manual annotation (Qamar et al., 2025). Therefore, in this study, we first used an LLM (GPT-4-turbo) to initially assign DAs, which were then manually corrected. This two-step approach was employed to maintain high accuracy while reducing the annotation workload compared to creating annotations from scratch.

3 Collaborative Story-Writing Dialogue

Our goal is to gain new insights into eliciting ideas from human–human co-creation, with the aim of contributing to the future design of human-AI co-creative systems. To this end, we assigned two human participants to different roles: a leader and a supporter. This asymmetric setup is designed to approximate a future human–AI co-creation environment, where humans lead the dialogue and generate ideas, and the system supports them. However, it is important to note that our focus is on analyzing genuine human–human dialogue; this is not a Wizard of Oz experiment where a human

simulates an AI.

3.1 Data Collection

Workers were recruited and matched using the crowdsourcing website Lancers [1]. They were asked to create an interesting story related to a given theme in Japanese. Each session involves two workers, divided into the roles of leader and supporter. The leader leads the dialogue and tackles the given creative task, specifically by thinking of interesting stories and writing them on the interface while discussing with the supporter. The supporter can only discuss with the leader via chat. Because this asymmetric setup is designed to approximate a future human–AI co-creation environment where a human leads and the system supports, we primarily analyze the supporter's dialogue strategy to elicit leader ideas in this study.

Workers are required to create at least one story of five to ten sentences on a given theme in 30 min. We built a story co-writing interface using Google Spreadsheets (see Appendix A). We used Google Spreadsheet because it enables the recording of worker operations and editing history through Google Apps Script. The theme is automatically generated by randomly assigning two characters and a genre to the template "(Character 1) and (Character 2) in (Genre)." For instance, the interface generates "merchants and demon lords in adventure story" or "futurists and old man in comedy." We used 52 types of characters and eight genres. The interface includes an input field where the leader worker enters the created story. Workers must create at least one story related to the given theme. They can create two or three stories simultaneously or create a second one after finishing the first one.

Data collection was performed using Zoom[2]. First, the leader worker shared the interface with the supporter using Zoom's screen sharing options. Second, the leader worker generated a theme on the interface and decided on the story's theme to be created. The leader worker was allowed to regenerate themes if they found a theme difficult to write stories about.

The co-creation dialogue began when the leader and the supporter agreed on a theme. The session was set to 30 minutes. Workers could only discuss via Zoom's text chat, as voice calls were prohibited.

[1]https://www.lancers.jp/

[2]https://www.zoom.com/

Postsurvey Item
You generated many ideas
You generated good ideas
You made pertinent points
Your partner generated many ideas
Your partner generated good ideas
Your partner made pertinent points
Your partner was easy to talk to
Your partner stimulated your idea generation
The conversation with your partner was lively
Overall, your partner was a good partner

Table 1: Items of the postsurvey

The leader worker could post to the chat and fill in the interface at any time. Workers could participate in the experiment as many times as they wanted. However, the same pair could not participate in the same roles.

Through a presurvey, we collected demographic data such as age and gender. In the postsurvey, participants used a seven-point Likert scale (Table 1) to answer questions regarding their self-evaluation and their evaluation of their partner. For example, "You generated many ideas" and "Your partner generated many ideas."

To answer RQ1, "In a co-creative dialogue, what is the relationship between dialogue acts and the quantity of ideas generated, as well as the evaluation of the partner?", we specifically analyze the postsurvey responses to the items concerning their own idea generation ("You generated many ideas") and their evaluation of the partner ("Overall, your partner was a good partner").

3.2 Data Collection Results

A total of 120 workers participated in our data collection. Initially, 500 dialogues were collected. Then, incomplete data were excluded. As a result, answers to a presurvey by 120 participants were obtained, as well as dialogue histories for 485 dialogues and answers to a postsurvey. A total of 118 workers participated as leaders, and 120 as supporters. Workers could take part in the experiment multiple times; on average, each worker participated 8.08 times (SD = 2.85), with a minimum of 2 and a maximum of 10 participations. Detailed demographic information for the workers is provided in Appendix C.

Figure 1 shows an example of the collected dialogues. The total number of utterances was 20,159

Theme

"Knight" and "princess" in human drama or youth story
(「騎士」と「姫」が出てくる「ヒューマンドラマ・青春物」)

Dialogue

Supporter: What kind of story are you thinking?
(どんな感じで作りますか？)
[setQuestion]

Leader: How about we make the main character a knight who's childhood friends with the princess?
(導入部分は主人公を騎士にして、姫と幼馴染的な展開にしますか？)
[suggest]

Supporter: The childhood friend idea sounds good.
(幼馴染いいですね。)
[positiveOpinion]

Leader: So maybe the princess starts falling for him without even realizing it after he protects her?
(姫は守ってもらったのをきっかけに知らずに知らずのうちに恋に落ちる的な？)
[suggest]

Supporter: I see.
(なるほど。)
[accept]
...

Created story

Once upon a time, there lived a princess and a knight who served in her castle.
(あるところにお姫様とそのお城に仕える騎士がいました。)

They were childhood friends and played together like brothers and sisters when they were young. (二人の関係は幼馴染で小さい頃は兄妹のように仲よく遊んで暮らしていました。)

As they grew older, they became more aware of their respective stations as a knight and a princess, and a distance began to grow between them.
(二人は成長しお互い騎士とお姫様という関係を実感し始め、いつしか距離を取り始めるようになりました。)
...

Figure 1: Example of dialogue during co-creation and a created story (originally written in Japanese, translated by authors). Dialogue acts (DAs) annotated are indicated in brackets.

and the total number of words was 287,077. Then, the average number of utterances per dialogue was 41.6 (SD = 16.6), and the average number of words was 591.9 (SD = 258.6).

3.3 Annotation

To analyze dialogues, we annotated the collected dialogues with dialogue acts (DAs). In this study, we designed new DAs specifically suited for co-creation dialogue.

We assigned provisional DAs to each sentence and repeated the modification of DAs. These tags were inspired by the tags in Hazumi (the multimodal dialogue corpus) (Komatani and Okada, 2021) and the ISO standard 24617–2 tags (Bunt et al., 2012). We decided to use 17 DAs as the final version, as shown in Appendix D.

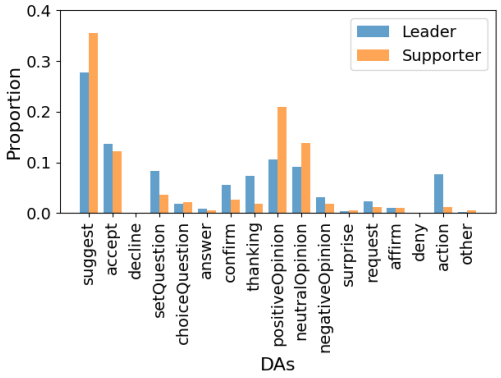

Figure 2: Distribution of DAs for supporters and leaders. The blue bars indicate the proportion of each tag relative to the total number of DAs for leaders, and the orange bars indicate the proportion for supporters.

All dialogues were annotated with GPT-4-turbo [3], followed by manual correction by human annotators. The GPT-4 prompt comprises three parts: the annotation manual, which includes tag definitions; a dialogue created by the authors; and the annotation results of the dialogue (see Appendix E). The Cohen's kappa value between the GPT-4-turbo annotations, which were manually corrected, and human annotations described in the previous paragraph was 0.78. For comparison, when two human workers annotated the five dialogues, the Cohen's kappa between them was 0.79. Therefore, this indicates that using GPT-4-turbo for initial annotation followed by manual correction achieves an accuracy level comparable to that obtained when employing manual annotation by human workers.

Figure 1 shows an example of annotation results. Tags are indicated within square brackets. Figure 2 demonstrates the distribution of the DAs of supporters and leaders. A comparison of leaders and supporters reveals that the proportion of leaders' *setQuestion* (question without options) is higher than that of supporters' *setQuestion*. This indicates that leaders seek more opinions from supporters. Additionally, the proportion of supporters' *positiveOpinion* (positive opinions or impressions) is higher than that of leaders' *positiveOpinion*. This indicates that the leader leads the dialogue, and the supporter responds to the leader's utterance, which is what we intended. Furthermore, the proportion of supporters' suggestions is higher than that of leaders' suggestions. This is because the leader directly inputs ideas into the interface.

[3] https://openai.com/

4 Analysis

Although this study analyzes human–human dialogue, the leader's role is intended to represent the user when co-creating with AI. We aim to analyze dialogues designed to elicit more ideas from leaders.

4.1 Postsurvey Analysis

To investigate the relationship between the leader's quantity of ideas and other factors, we analyzed the results of the postsurvey using logistic regression. In this analysis, as shown in Table 1, the leader's response to "You generated many ideas" was used as the dependent variable, with the other items serving as independent variables. However, to avoid issues of multicollinearity, "Overall, your partner was a good partner" was excluded from the model. Furthermore, the values of the dependent variable were binarized: values greater than or equal to the median were converted to 1, and values below the median were converted to 0. Table 2 shows the results of the logistic regression analysis. Focusing on the statistically significant results, the quality of the leader's own ideas ("You generated good ideas") and the pertinence of their points ("You made pertinent points") had a positive association with the dependent variable. This indicates that when leaders generate a high quantity of ideas, they also tend to generate high-quality ideas and make pertinent points. In contrast, the quantity of the partner's ideas ("Your partner generated many ideas") had a negative association with the dependent variable. This suggests that when the leader's quantity of ideas is high, the supporter's quantity of ideas tends to be low. Possible reasons for this include a leader subjectively feeling their own contribution is smaller when their partner's is larger, or feeling a diminished need to generate ideas themselves when their partner is highly generative.

4.2 Relationship between Evaluation Metrics and Dialogue Acts

The objective of this study is to identify the characteristics of dialogue that facilitate a leader's idea generation. To this end, we analyze the relationship between dialogue acts and the three evaluation metrics detailed below.

Quantity of ideas (subjective) The first metric subjectively evaluates the quantity of ideas generated by the leader. The leader's response score to "You generated many ideas" in the postsurvey was used as the score.

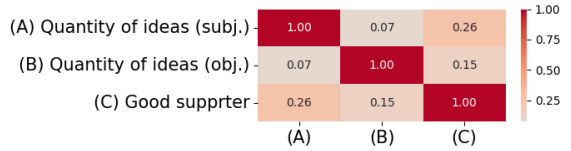

Figure 3: Correlation among three evaluation criteria (Quantity of ideas (subjective) / Quantity of ideas (objective) / Good supporter)

Quantity of ideas (objective) The second metric objectively evaluates the quantity of ideas generated by the leader. The number of *suggest* tags in the leader's utterances is used as the score. *Suggest* tag is assigned to utterances proposing ideas or action. This criterion is employed for an objective evaluation because the first criterion is subjective and there might be a discrepancy between the actual quantity of ideas and subjective evaluation.

Good supporter The third metric evaluates the supporter's positive influence on the leader. When collaborating with a dialogue system, assessing whether it was a good collaborative partner will be anticipated; therefore, we adopted this criterion in this study. The leader's response score to "Overall, your partner was a good partner" in the postsurvey was used as the score for "Good supporter." This criterion correlated with partner-related questions, which indicates the overall evaluation of the supporter, making it a suitable evaluation criterion (see Appendix B).

We conducted a Spearman rank correlation analysis among the evaluation metrics. As shown in Figure 3, the resulting correlations were generally weak. Furthermore, no significant correlation was found between the two metrics for idea quantity, "Quantity of ideas (subjective)" and "Quantity of ideas (objective)" ($\rho = 0.07, p > 0.05$). This result suggests the importance of measuring the quantity of ideas from both subjective and objective perspectives.

Figure 4 shows the results of the correlation analysis between the evaluation metrics and the number of the leader's and supporter's dialogue acts. For this analysis, we used different correlation methods based on the data type of the evaluation metric. For the relationship with "Quantity of ideas (objective)," we used the Pearson correlation coefficients. For

	Coefficient (β)	Std. Error (SE)	z-value	p-value
Intercept	−15.97	1.61	−9.89	***
You generated good ideas	2.49	0.26	9.45	***
You made pertinent points	0.43	0.17	2.54	*
Your partner generated many ideas	−0.66	0.20	−3.23	**
Your partner generated good ideas	−0.08	0.27	−0.28	
Your partner made pertinent points	0.34	0.22	1.53	
Your partner was easy to talk to	0.39	0.24	1.62	
Your partner stimulated your idea generation	0.03	0.23	0.15	
The conversation with your partner was lively	0.09	0.23	0.42	

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Table 2: The results of the logistic regression analysis with the leader's response to "You generated many ideas" as the dependent variable.

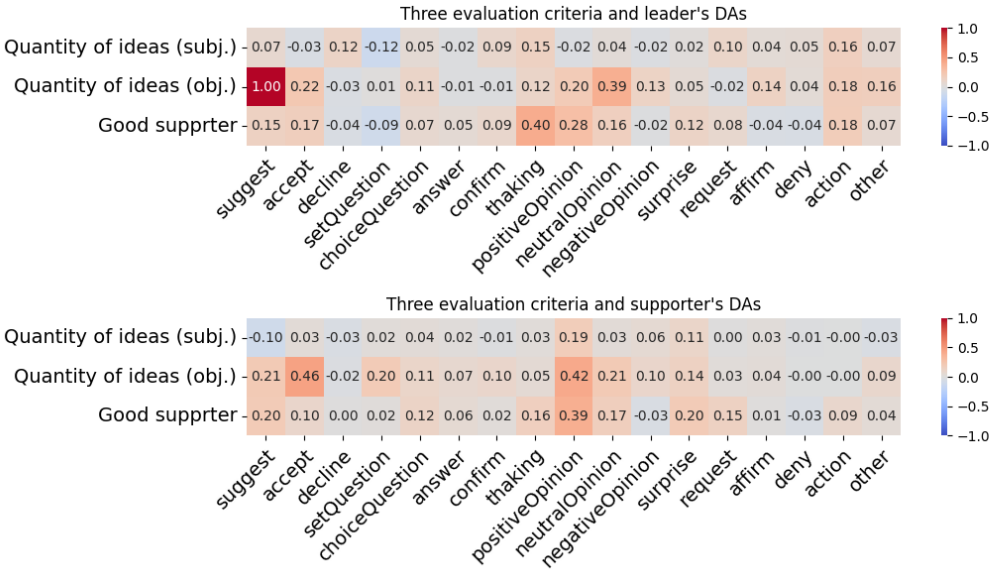

Figure 4: Correlation between three evaluation criteria (Quantity of ideas (subjective) / Quantity of ideas (objective) / Good supporter) and DAs

the relationships with "Quantity of ideas (subjective)" and "Good supporter," which are based on ordinal Likert scales, we used the Spearman rank correlation coefficients.

As shown in the top panel of Figure 4, there were weak to moderate correlations between the "Quantity of ideas" metrics and the leader's dialogue acts. Regarding the "Good supporter" metric, an interesting point is the positive correlation observed with *thanking* ($\rho = 0.40, p < 0.01$). *Thanking* is a DA that expresses gratitude, such as "Thank you." We discuss this result in detail in Section 4.3.

Then, as shown in the bottom panel of Figure 4, "Quantity of ideas (objective)" had weak to moderate correlations with the supporter's DAs. Notably, it showed a moderate positive correlation with the supporter's *accept* ($\rho = 0.46, p < 0.01$). This suggests that an increase in acceptance in the supporter's utterances could potentially lead to an increase in proposals from the leader. A moderate correlation was also found between "Quantity of ideas (objective)" and the supporter's *positiveOpinion*. This indicates that when the leader makes many proposals, the supporter tends to respond with positive feedback, such as "that's good," which can be interpreted as a reaction to the leader's high number of proposals. Furthermore, there was a moderate positive correlation between the "Good supporter" metric and the supporter's *positiveOpinion* ($\rho = 0.39, p < 0.01$). This result suggests that receiving positive opinions from a partner may lead to a more favorable overall impression of that partner. Given that the supporter's *positiveOpinion* is positively correlated with both subjective and objective measures of idea quantity, as well as with the partner evaluation, it can be considered a particularly useful dialogue act when designing strategies to elicit ideas.

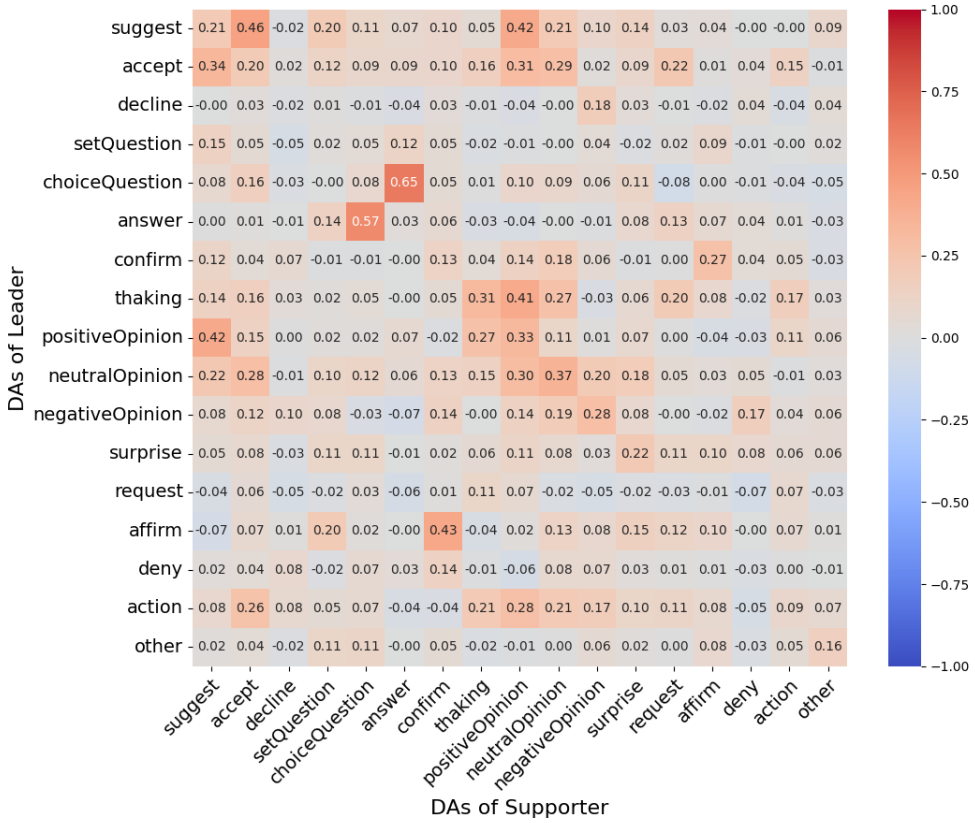

Figure 5: Pearson correlations between leaders' DAs and supporters' DAs

4.3 Analysis of Leader's and Supporter's Dialogue Acts

To analyze the relationship between the dialogue acts of the leader and the supporter, we performed a correlation analysis using the Pearson correlation coefficient.

As shown in Figure 5, some pairs of dialogue acts exhibited correlations that are expected by their definitions. For instance, a strong positive correlation was found between *choiceQuestion* (a question presenting two or more options, e.g., "Is the character female or male?") and *answer* (a response to a *choiceQuestion*, e.g., "Female."). Similarly, a moderate positive correlation was observed between *confirm* (an utterance seeking confirmation) and *affirm* (an affirmative response). These results are consistent with the definitions of the dialogue acts. Furthermore, this analysis revealed a positive correlation between the supporter's *positiveOpinion* and the leader's *thanking*. This suggests a conversational pattern where a positive utterance from the supporter is often followed by an expression of gratitude from the leader. Considering this in conjunction with the finding from Figure 4 that *thanking* correlates positively with the "Good supporter" metric, a plausible causal chain emerges. It is likely that the supporter's use of *positiveOpinion* contributes to the leader's perception of them as a "Good supporter," while also eliciting *thanking* from the leader. Therefore, the observed correlation between *thanking* and "Good supporter" may be an indirect consequence.

5 Discussion

5.1 RQ1: What is the relationship between dialogue acts and the quantity of ideas generated, as well as the evaluation of the partner?

As shown in Figure 3, no correlation was found between the two metrics for idea quantity, "Quantity of ideas (subjective)" and "Quantity of ideas (objective)." This discrepancy does not suggest contradictory results but highlights the potential disparity between subjective and objective evaluations of the quantity of ideas. For example, even if the quantity of the leader's ideas was quantitatively high, the evaluation might be relative to the quantity of the supporter's ideas.

In this study, we analyzed human–human dialogue to identify characteristics that elicit ideas. However, a limitation is that the relationships observed between dialogue acts, the quantity of ideas, and partner evaluations were generally weak to moderate, with no strong correlations found. This suggests that a complex creative process like idea generation is likely influenced by a variety of factors beyond the dialogue acts.

5.2 RQ2: What is the relationship between one's own dialogue acts and the partner's dialogue acts?

Our goal is to design dialogues that elicit more ideas from leaders. Therefore, we discuss the supporter's dialogue acts that were positively correlated with the objective idea quantity (Figure 5). *Accept* refers to utterances that accept a proposal. The tendency for the supporter's *accept* to be high when the leader's *suggest* is high can be attributed to the supporter actively accepting the leader's proposals. However, the direction of causality in these correlations could be reversed. For instance, it is possible that a supporter's frequent use of *accept* fosters an environment where the leader feels more comfortable making suggestions. Similarly, it is unclear whether a positive atmosphere created by the supporter fosters more ideas from the leader, or if the supporter's positive reactions are simply a consequence of the leader being highly generative.

5.3 Human–AI Co-creation Design

This study aimed to identify the features of utterances that facilitate idea generation by analyzing dialogue data from a human–human co-creative process. In this section, we discuss dialogue strategies for eliciting human ideas in human–AI co-creation based on our findings. Our results revealed that the quantity of the supporter's ideas had a negative association with the leader's quantity of ideas, suggesting that receiving fewer ideas from a partner may, in fact, stimulate one's own ideation. However, it is crucial to note that this finding is limited to the subjective measure of idea quantity; no negative correlation was found for the objective measure (i.e., between the number of *suggest* acts from leaders and supporters). This suggests the negative relationship may only apply when idea quantity is perceived subjectively. We also found that acceptance and positive opinions from supporters were positively correlated with the leader's idea proposals, indicating that these can be effective supportive strategies. These findings offer important design implications for future co-creative AI systems. For instance, instead of always presenting numerous ideas, an AI could adopt a strategy of accepting the user's proposals and providing positive feedback to better draw out their creativity. Furthermore, we found a positive correlation between the positive opinions of the supporter and the leader's overall evaluation of that partner. This suggests that positive feedback may not only increase the quantity of ideas but also enhance the user's impression of their collaborative partner. However, since our results are based on human-human interaction, whether the same outcome would occur with an AI partner requires future validation. Future work includes examining the causality of the correlations identified in this study and building and evaluating a dialogue system that implements these findings.

6 Conclusion

In this study, we investigated utterance patterns that promote idea generation in human–human co-creative dialogue, aiming to inform the design of AI systems that can better elicit human creativity. We conducted a collaborative story-writing experiment with pairs assigned leader and supporter roles and analyzed the collected dialogue data. Our results revealed the leader's self-assessed idea quantity decreased as their perception of the supporter's idea quantity increased. We also found that acceptance and positive opinions from supporters were positively correlated with the leader's idea proposals, indicating that these can be effective supportive strategies.

Limitations

This study has several limitations. First, our study's participants were limited to Japanese native speakers. This demographic specificity may restrict the generalizability of our findings.

Second, this study focused on human–human dialogues rather than human–AI interactions. While our ultimate goal is to inform the design of human–AI co-creative systems, we opted for human–human data at this exploratory stage. Building on the findings of this paper, our future work will apply these insights to conduct controlled experiments in human–AI co-creative settings.

Third, as shown in Figure 4, the primary relationships observed were generally weak to moderate correlations, with no strong correlations found.

This may suggest that a complex process like idea generation is influenced by many factors beyond dialogue acts. Future research should consider incorporating factors such as the leader's personality and their proficiency in story writing into the analysis.

Fourth, while we observed a positive correlation between a supporter's DAs and the "Quantity of ideas (objective)," the causal direction of this relationship remains undetermined. It is unclear whether a high frequency of a particular DA from the supporter actively stimulated the leader's idea generation, or if the supporter's DAs (such as acceptance) simply increased in response to the leader proposing a large number of ideas. As part of our future work, establishing this causality will require further experiments designed specifically to distinguish between these two possibilities.

References

Minwook Bae and Hyounghun Kim. 2024. Collective Critics for Creative Story Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819, Miami, Florida, USA. Association for Computational Linguistics.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 196–207.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J. Cai. 2020. AI Song Contest: Human-AI Co-Creation in Songwriting. In *International Society for Music Information Retrieval (ISMIR)*.

Gloria Ashiya Katuka, Alexander R. Webber, Joseph B. Wiggins, Kristy Elizabeth Boyer, Brian Magerko, Tom McKlin, and Jason Freeman. 2022. The Relationship between Co-Creative Dialogue and High School Learners' Satisfaction with their Collaborator in Computational Music Remixing. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1).

Mika Koivisto and Simone Grassini. 2024. Author Correction: Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 14.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. SceneCraft: Automating Interactive Narrative Scene Generation in Digital Games with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 19(1):86–96.

Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Alex Faickney Osborn. 1953. *Applied imagination.* New York: Charles Scribner's Sons.

Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 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*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Ayesha Qamar, Jonathan Tong, and Ruihong Huang. 2025. Do LLMs Understand Dialogues? A Case Study on Dialogue Acts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237, Vienna, Austria. Association for Computational Linguistics.

Peinuan Qin, Chi-Lan Yang, Jingshu Li, Jing Wen, and Yi-Chieh Lee. 2025. Timing Matters: How Using LLMs at Different Timings Influences Writers' Perceptions and Ideation Outcomes in AI-Assisted Ideation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Oliver Schmitt and Daniel Buschek. 2021. Character-Chat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Proceedings of the 13th Conference on Creativity and Cognition*, C&C '21, New York, NY, USA. Association for Computing Machinery.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pages 841–852.

A Interface for Co-creation

Figure 6 shows the story co-writing interface. The story co-writing interface was built on Google Spreadsheets. Only the leader could write the story on the interface. When a worker clicks on the theme generation button on the interface, it automatically generates and shows the theme of the story being created. The button for automatic theme generation is placed on another sheet to prevent accidental touches. The generated theme is displayed in the yellow area of the interface. Participants create their story based on this theme, entering it sentence by sentence into the designated cells. The input field consists of ten fields per story and a checkbox indicating that the story creation is complete.

B Relationship between the three evaluation criteria and the leader's responses in the postsurvey

Table 3 presents the Spearman rank correlation coefficients between the three evaluation criteria and leader's postsurvey responses. There was a strong correlation between the "Quantity of ideas (subjective)" and generating good ideas. One of Osborn's principles of brainstorming is "focus on quantity, more ideas will increase the likelihood of good ideas" (Osborn, 1953). The obtained result supports the aforementioned Osborn's brainstorming principle. We confirmed that "Quantity of ideas (subjective)" is suitable for evaluating the quality and quantity of the leader's ideas.

Additionally, there were strong correlations between "Good supporter" and each response to Questions 4–9 in the survey. This implied the overall evaluation of the supporter, making the criterion a suitable evaluation criterion.

C Participants

In this section, the demographic information of the 120 workers who participated in this experiment is provided. Of the participants, 60.8% were female and 39.2% were male. Additionally, 7.5% were in their teens or younger, 35.8% were in their 20s, 26.7% were in their 30s, 20.8% were in their 40s, 7.5% were in their 50s, and 1.7% were in their 60s.

D Definition of DAs

Table 4 shows the definition of 17 DAs which were used for annotations.

E Prompt for annotation

Initial dialogue act annotation was performed using GPT-4-turbo. The prompt for annotation consisted of three parts (Figure 7): the annotation manual (task description, tag list, and notes of annotation); a dialogue example created by the authors (input example 1); and the corresponding annotation results for that dialogue (output example 1). The dialogue that we wanted to be annotated was appended to this prompt and then input into the model.

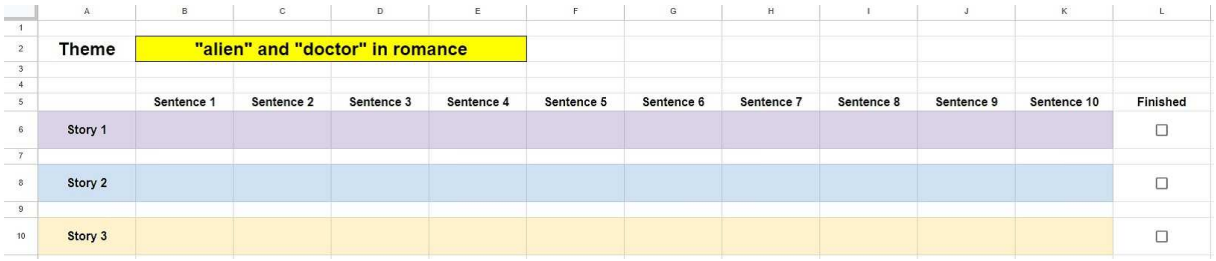

Figure 6: The story co-writing interface was built on Google Spreadsheets. The original interface is written in Japanese. The button for automatic theme generation is placed on another sheet to prevent accidental touches.The yellow area indicates the story's theme. The first story is entered sentence by sentence into the cells of the purple area, the second story into the light blue area, and the third story into the light yellow area.

No.	Postsurvey Item	Subjective Eval.		Objective Eval.
		Good supporter	**Quantity of ideas (subjective)**	**Quantity of ideas (objective)**
Self Evaluation				
1	You generated many ideas	0.26	-	0.07
2	You generated good ideas	0.27	0.78	0.02
3	You made pertinent points	0.10	0.53	−0.04
Partner Evaluation				
4	Your partner generated many ideas	0.56	0.23	0.07
5	Your partner generated good ideas	0.74	0.27	0.12
6	Your partner made pertinent points	0.68	0.31	0.10
7	Your partner was easy to talk to	0.85	0.35	0.13
8	Your partner stimulated your idea generation	0.80	0.32	0.12
9	The conversation with your partner was lively	0.81	0.37	0.11
10	Overall, your partner was a good partner	-	0.26	0.15

Table 3: Relationship between the three evaluation criteria and other responses in the postsurvey

Table 4: The definition of DAs. The underlined part represents the utterance to which the tag is annotated.

Tag Name	Tag Description	Utterance Example
suggest	Proposal of ideas or directions. It is a statement or question sentence.	Let us introduce pirates.
accept	Acceptance of proposals or agreement with the partner's opinion.	It's getting better, isn't it? → Yes.
decline	Rejection of proposals or disagreement with the partner's opinion.	It's getting better, isn't it? → No.
setQuestion	Question without options.	Where is a good setting?
choiceQuestion	Question presenting two or more options.	Are idols female? Male?
answer	Response to a choiceQuestion.	Are idols female? Male? → Male.
confirm	Confirmation of ideas or facts.	Is this okay?
thanking	Expressing gratitude.	Thank you.
positiveOpinion	Positive opinions or impressions.	That's good!
neutralOpinion	Opinions or impressions that are neither positive nor negative.	Punctuation is missing.
negativeOpinion	Negative opinions or impressions.	It doesn't add up.
surprise	Expressing surprise.	Oh!
request	Request to the partner.	Please read.
affirm	Affirmation regarding facts, with no intention of agreeing.	Does it mean like this? → Yes.
deny	Denial regarding facts, with no intention of disagreeing.	Does it mean like this? → No.
action	Declaration of future actions.	I will write it.
other	Anything that does not fit into the above tags.	It's been about 15 min.

Task Description

- This is a task to annotate text dialogue data between two people.
- The content of the dialogue is that two people who are given a theme create a story that matches the theme while brainstorming ideas.
- Please annotate considering the entire context, not just a single utterance.

Tag List

[Definition of DAs]

Notes on Annotation

…

- Output Format: Utterance<Reason for assigning the tag>[Assigned tag]
- Please assign only one tag per utterance. If there are two or more candidate tags, please select the one you think is best.

Input Example 1

…

A: Should the meeting of the princess and the pirate be on a ship?

A: Or in a castle?

B: Let's have it on a ship.

…

Output Example 1

…

A: Should the meeting of the princess and the pirate be on a ship? <Reason: When the same speaker presents choices in two separate utterances, apply the “choiceQuestion" tag to both.>[choiceQuestion]

A: Or in a castle? <Reason: When the same speaker presents choices in two separate utterances, apply the “choiceQuestion" tag to both.>[choiceQuestion]

B: Let's have it on a ship<Reason: Answering a question with choices.>[answer]

…

#Input Example 2

[Dialogue history]

#Ourput Example 2

Figure 7: Prompt for DA annotation. The original prompt is written in Japanese.