StoryCCDial: Collecting and Analyzing Human–Human Co-Creation Dialogues for Personalized Creative Support

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

Abstract

With the development of generative models, research on human–AI co-creation has been actively conducted. However, in the field of co-creation, research on system personalization according to individual characteristics is insufficient, and little focus has been placed on individual differences in creation. Therefore, in this study, we constructed StoryCCDial, a co-creation dialogue dataset aimed at the personalization of co-creative dialogue systems. First, we collected human–human story co-creation dialogue data involving 120 workers and constructed a dataset that includes dialogues, dialogue acts, the workers' personality traits, postsurveys, and edit histories from the interface. Next, using the constructed dataset, we conducted analyses focusing on the workers' personality traits, the number of utterances, and edit histories. The analysis revealed differences in dialogue content based on workers' personality traits, individual differences in the number of utterances during the co-creation process, and variations in creative workflows on the interface. Our dataset will be available at https://github.com/UEC-InabaLab/StoryCCDial.

Keywords: Data collection, Co-creation, Dialogue

1. Introduction

Research on computer-aided creative support is actively being conducted. These studies focus on theater scripts (Mirowski et al., 2023), cartoon character design (Hiruta et al., 2022), persuasive texts (Munigala et al., 2018), slogans (Alnajjar and Toivonen, 2021; Kim et al., 2023), novels (Yuan et al., 2022), and lyrics parodies (Gatti et al., 2017). With the development of generative models, human–artificial intelligence (AI) co-creation systems have been widely studied. In these systems, users interact with AI to collaboratively work on creative tasks, issuing commands for partial modifications or 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; Yuan et al., 2022). Human–AI co-creation systems that interact with people through natural language have also been investigated (Schmitt and Buschek, 2021). However, research focusing on dialogue during such co-creation is still limited, and knowledge regarding suitable dialogue during co-creation remains insufficient. In particular, to the best of our knowledge, no research has focused on dialogue adapted to each individual user in co-creation. In this research, with the aim of building a dialogue system that provides dialogue suitable for a user's characteristics in co-creation, we collected and will release a dataset, StoryCCDial (Story Co-Creation Dialogue). This dataset includes dialogues, dialogue acts, edit histories, the workers' personality traits, and postsurveys.

The task involves two people assigned the asymmetric roles of "leader" and "supporter," co-creating a story. The data includes the workers' personalities, dialogue data, postsurvey data on their partner and themselves, and interface edit histories. Notably, we annotated the dialogue data with dialogue act tags and analyzed the relationship between individual characteristics and the content of the dialogue. We also analyzed variations within the co-creation dialogues by examining the number of utterances and edit histories.

The contributions of this study are as follows:

- We collected and will release a human–human co-creation dialogue dataset. This dataset includes dialogues, dialogue acts, edit histories, the workers' personality traits, and postsurveys.
- We analyzed co-creation dialogue by focusing on individual characteristics (personality traits and the number of utterances) and characteristics of each co-creation workflow (edit histories). As a result, this suggested that the dataset includes diverse workers and can be utilized for personalized systems.

It should be noted that a previous study (Ezure and Inaba, 2025) has already analyzed the relationship between dialogue acts and the quantity of ideas in the co-creation dialogue within this dataset. Consequently, this paper does not address this topic.

2. Related Work

2.1. Co-creation

Several studies have been conducted on human–AI co-creation systems using generative models

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1934–1946
11–16 May 2026.

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 keywords. Dramatron (Mirowski et al., 2023) is a collaborative scriptwriting system built using Chinchilla (Hoffmann et al., 2022). From the input log line, Dramatron can generate an entire script with a title, list of characters, a plot, location descriptions, and dialogue. 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 a leader incrementally refine drafts of the plan and story over multiple rounds.

CharacterChat (Schmitt and Buschek, 2021) is designed for users to create fictional characters while chatting with the system. Similar to CharacterChat, our work focuses on building a story through dialogue. However, our study distinguishes itself by analyzing human–human collaborative interactions for AI system design.

Zhou et al. (2024) analyzed dialogue data and edit histories for human–human co-creation, which is close to our research. However, our study differs in that it aims at the personalization of co-creation.

2.2. Personality-aware Systems and Analysis

Personality traits influence human decision-making processes, user preferences, and interests (Rentfrow and Gosling, 2003). In the field of psychology, several studies have been conducted from various perspectives on the effects of personality, such as political choices (Caprara et al., 2006), counterproductive behaviors (Salgado, 2002), and creativity (Feist, 1998).

In the field of computer science, personality-aware recommendation systems have attracted significant attention, including recommendations of music (Perik et al., 2004; Hu and Pu, 2010; Liu and Hu, 2020), movies (Karumur et al., 2018; Nguyen et al., 2018; Nalmpantis and Tjortjis, 2017), places of interest (Elahi et al., 2013; Tkalcic et al., 2009; Braunhofer et al., 2015), games (Yang and Huang, 2019; Zeigler-Hill and Monica, 2015), news (Dhelim et al., 2020), pictures (Li et al., 2019, 2020; Gelli et al., 2017), and friends on social networking services (Neehal and Mottalib, 2019; Ning et al., 2019). Personality-aware recommendation systems are effective in solving the cold start, free-rider, and data sparsity problems, which are issues in conventional recommendation models. Fatahi et al. (2023) investigated what kind of persuasive messages should be presented in music recommendations based on user personality traits.

Thus, the effectiveness of personalization based on personality traits has been suggested in the field of recommendation systems. In this study, we conduct an analysis focusing on personality traits to explore the potential for personalization in the different domain of co-creation. Particularly, based on previous research indicating a link between creativity and Openness (McCrae, 1987), we focus our analysis on this personality trait to examine its effects in detail.

3. Collection of Co-Creation Dialogue

In this study, we collected human–human Story Co-Creation Dialogue (StoryCCDial) dataset to develop a personalized co-creation dialogue system.

3.1. Story Co-Creation Task

The workers co-created a story while brainstorming ideas. They were instructed to create an interesting story on a given theme in Japanese. In each session, two workers were paired up and assigned the roles of “leader” and “supporter.” The leader was responsible for sharing the interface screen and writing the story on the interface through discussion with the supporter. On the other hand, the supporter only communicated through text chat. The reason for this asymmetric role design is that it simulates a human–AI co-creation environment where a human leads and a system supports. The workers were asked to create at least one story on the given theme, consisting of 5 to 10 sentences, within 30 minutes.

3.2. Interface

We built the interface for writing the story using Google Spreadsheets (Figure 1). The reason for using Google Spreadsheets is that it allows us to record the workers’ edits through Google Apps Script. The theme is automatically and randomly generated with two characters and a genre along the template “(Character 1) and (Character 2) in (Genre).” For example, the interface generates themes such as “merchants and demon lords in adventure story” or “a futurist and an 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. However, only one story was created in over 95% of the collected data. Therefore, this

study focuses on analyzing the first story and its related postsurveys. Data collection was performed using Zoom[1].

The edit history is stored as a series of logs, where each entry records a timestamp alongside one of three actions: writing text for the story, deleting text from the story, or checking a checkbox to indicate whether the story is complete.

3.3. Data Collection Procedure

Workers were recruited and matched using the crowdsourcing website Lancers[2]. First, workers answered a presurvey. Through the presurvey, we collected demographic data, such as age and gender, as well as personality traits. To measure their personality traits, we used a 10-item questionnaire based on the Japanese version of the Ten Item Personality Inventory (TIPI-J) (Oshio et al., 2012) to assess the workers' Big Five personality traits. Second, the leader shared the interface with the supporter using Zoom's screen sharing option. Third, the leader generated a theme on the interface to decide on the theme of the story to be created. If they found the theme difficult to create a story with, the leader was allowed to regenerate the theme. The co-creation dialogue began once the pair agreed on a theme. Fourth, with a time limit of 30 minutes per session, they collaboratively created an interesting story based on the theme while exchanging ideas. The workers could only discuss through Zoom's text chat; voice calls were prohibited. The leader could chat and write on the interface at any time. Workers were allowed to participate in the experiment as many times as they wished. However, they were not allowed to participate with the same partner in the same role. Fifth, after the co-creation dialogue, they answered a postsurvey. In the postsurvey, workers evaluated themselves, their partners, and the created story using a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). For example, questions included items such as "You generated many ideas," "Overall, your partner was a good partner," and "The story is interesting overall." A list of the postsurvey items is shown in Table 2.

3.4. Data Statistics

A total of 120 workers participated in our data collection. Initially, 500 dialogues were collected. Then, incomplete data were excluded. In the result, answers to a presurvey by 120 participants were obtained, as well as dialogue histories for 485 dialogues and answers to a postsurvey. Statistical information on the collected data is shown in Table 1. The workers were 39.2% male and 60.8% female, with the largest number of participants in their 20s. 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 1 and a maximum of 10 participations.

Category	Value
Number of participants	120
Gender	
Male	39.2%
Female	60.8%
Age	
19 or younger	7.5%
20–29	35.8%
30–39	26.7%
40–49	20.8%
50–59	7.5%
60–69	1.7%
Number of dialogue histories	485
Avg. utterances per dialogue history	41.6
Avg. words per dialogue history	591.1
Number of completed stories	497
Avg. number of sentences in completed stories	10.0
Avg. number of words per sentence in completed stories	34.6
Number of edit histories	480
Avg. number of edit actions per edit history	23.1

Table 1: Statistics of the participant demographics and the collected dialogue corpus.

Based on the responses to the presurvey, we measured a score of 2–14 for each of the Big Five personality traits using TIPI-J. Figure 2 shows the distribution of the participants' Big Five personality traits. We found that most participants had high scores on agreeableness and openness personality traits.

Figure 3 shows an example of the collected dialogues. The total number of utterances was 20,159 and the total number of words was 286,687. Then, the average number of utterances per dialogue was 41.6 (SD = 16.6), and the average number of words was 591.1 (SD = 258.2).

Table 2 shows the results of the postsurvey. Note that the response scores used in the table are from the leader's postsurvey. In this study, we assume that the leader is the user and the supporter is the system; therefore, for the postsurvey, our analysis focuses on the leader's responses. With possible values from 1 to 7, the means are all above 5 in Table 2, indicating that the overall evaluations in the postsurvey were high. Furthermore, "Overall,

[1] https://www.zoom.com/
[2] https://www.lancers.jp/

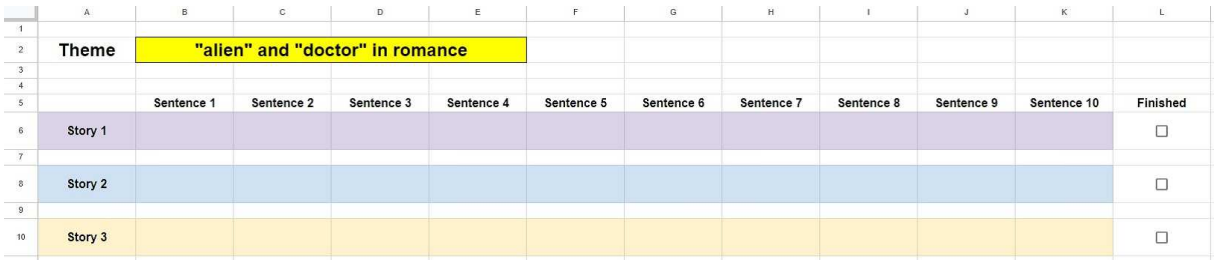

Figure 1: Story co-writing interface built on Google Spreadsheets. The original interface is written in Japanese. The button for automatic theme generation is placed on another sheet to prevent accidental touches. The yellow area indicates the story's theme. The first story is entered sentence by sentence into the cells of the purple area, the second story into the light blue area, and the third story into the light yellow area.

your partner was a good partner," was the highest of all postsurvey items at 6.04, indicating that leaders generally had a favorable impression of their partners.

Table 1 shows that there are 497 completed stories. A completed story is defined as one whose edit history includes checking a checkbox to indicate the completion of the story, or one that the authors judged to be complete even if the checkbox was left unchecked. Due to the nature of the task, which requires creating at least one story per dialogue, the number of completed stories (497) exceeds the total number of dialogues (485). Additionally, Table 1 shows there were 480 edit histories after incomplete data were excluded. Incomplete data refers to edit histories containing a high frequency of improperly recorded operations or operations performed outside the designated interface boundaries. Furthermore, the average number of edit actions per edit history was 23.1, the average number of sentences in a completed story was 10.0, and the average number of words per sentence was 34.6.

Item	Mean (SD)
Self-Evaluation	
You generated many ideas	5.74 (0.99)
You generated good ideas	5.70 (0.99)
You made pertinent points	5.14 (1.13)
Evaluation of Partner	
Your partner generated many ideas	5.78 (1.17)
Your partner generated good ideas	5.93 (1.17)
Your partner made pertinent points	5.89 (1.17)
Your partner was easy to talk to	5.98 (1.31)
Your partner stimulated your idea generation	5.84 (1.33)
The conversation with your partner was lively	5.68 (1.32)
Overall, your partner was a good partner	6.04 (1.28)
Evaluation of Generated Story	
The story is easy to understand	6.02 (0.86)
The story is consistent	6.00 (0.88)
The story is full of surprises	5.42 (1.26)
The story is interesting overall	5.70 (1.04)

Table 2: Results of postsurvey (7-point Likert scale). SD denotes standard deviation.

3.5. Annotation

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

First, one author randomly selected nine dialogues, segmented utterances in each dialogue into sentences, and assigned provisional DAs to each sentence. We analyzed the trends of the assigned DAs, merged them, and proposed 19 tags. 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 then discussed the proposed tags and tentatively determined tags by further splitting and merging. Two different workers were then asked to annotate the same five dialogues using the tentatively determined tags. The discrepancies were analyzed, and the tag types and definitions were integrated and modified. The same two workers were asked again to annotate the five new dialogues with the modified tags. As a result, the Cohen's kappa value for the annotations by the two workers reached 0.79, indicating sufficient agreement. Therefore, we decided to use the 17 tags used at that time as the final version, as shown in Table 3.

All dialogues were annotated with GPT-4-turbo [3], followed by manual correction by human annotators. Specifically, through this manual correction, appropriate dialogue acts were assigned to utterances that GPT-4-turbo frequently categorized as "other." The GPT-4-turbo prompt comprises three

[3] https://openai.com/

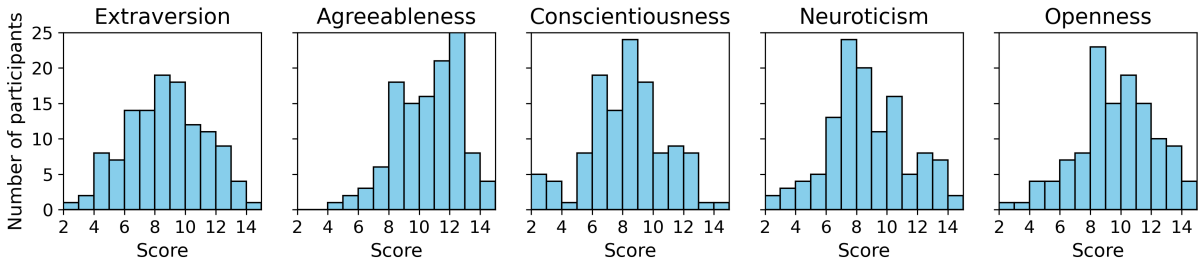

Figure 2: Distribution of the participants' Big Five personality traits

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.

Table 3: The definition of DAs. The underlined part represents the utterance to which the tag is annotated.

parts: the annotation manual, which includes tag definitions; a dialogue created by the authors; the annotation results of the dialogue (see Appendix A). 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 with that obtained when employing manual annotation by human workers.

Figure 3 shows an example of annotation results. Tags are indicated within square brackets. Figure 4 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

Theme

"God" and "merchant" in human drama or youth story
(「神様」と「商人」が出てくる「ヒューマンドラマ・青春物」)

Dialogue & Edit History

...
20:57:12 **Supporter:**
I think the idea of miracles happening one after another is an interesting development! (次々と奇跡が起こるって展開は面白いと思います！)
[positiveOpinion]

20:57:59 **Leader:**
Thank you. (ありがとうございます。)
[thanking]
I'm already stumbling on the first sentence. (最初の一文目からつまずきます。)
[negativeOpinion]
Is "a merchant who is terrible at business" a good way to start? (商売下手な商人って感じでいいですかね)
[suggest]

20:58:50 **Leader:**
write sentence1

20:59:02 **Supporter:**
That's great! (いいですね！)
[positiveOpinion]
Something like, he honestly tells his customers about the product's flaws and ends up selling nothing at all might be good! (よくない商品とかも正直にお客に教えてしまって全く売れないみたいな感じがいいかもですね！)
[suggest]
...

Created story

sentence1
In a certain world, there lived a merchant who was terrible at business. (とある世界に商売下手な商人が暮らしていた。)

sentence2
Although he could have sold his goods well,
he would foolishly and honestly tell his customers
about his products' flaws, and days went by
when he sold nothing at all.
(二人の関係は上手く売ればいいものの、
良くない商品を馬鹿正直にお客に伝えてしまい、
全く売れない日が続いていた。)

sentence3
Meanwhile, a mean-spirited merchant, who was a skilled liar, seemed to be thriving in his business by cleverly deceiving his customers. (一方で、嘘が得意のいじわる商人は、
うまい事客を騙して商売が繁盛しているようだ。)
...

Figure 3: 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.

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

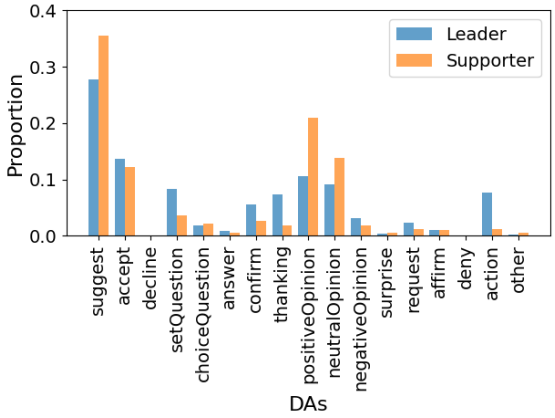

Figure 4: Distribution of DAs for supporters and leaders. The blue bars indicate the proportion of each tag relative to the total number of DAs for leaders, and the orange bars indicate the proportion for supporters.

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.

4. Analysis

The purpose of this study is data collection for building a personalized co-creation dialogue system. Therefore, for the collected data, we first analyzed the relationship between the leader's DAs and personality. The reason for analyzing the relationship with the leader is that, although this study analyzes human–human dialogue, the leader's role is intended to represent the user when co-creating with AI. Furthermore, to gain deeper insights into their behavior during co-creation, we also analyzed the number of utterances and conducted the clustering of edit histories.

4.1. Relationship between Leader's Personality Traits and Dialogue Acts

We analyzed the relationship between the leader's personality traits and their dialogue features. In this analysis, we used openness as the leader's personality trait. This is because prior research has shown that "the creativity is particularly related to the personality domain of openness to experience" (McCrae, 1987). We defined workers with an openness score of 8 or less as low-openness and those with a score of 9 or more as high-openness. Figure 5 shows the average number of occurrences of all 17 types of the leader's dialogue acts in a single dialogue. It was found that high-openness leaders had more *suggest* acts than low-openness leaders.

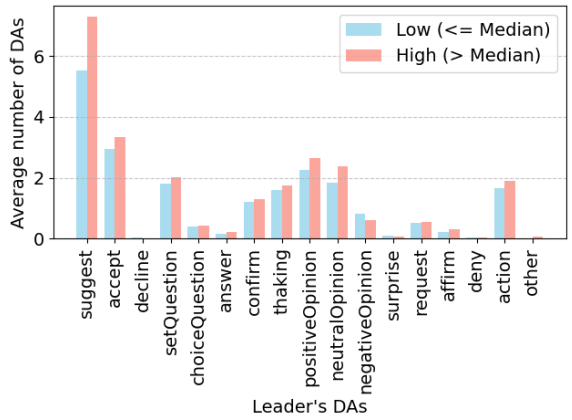

Figure 5: Average number of dialogue acts for low-openness and high-openness groups.

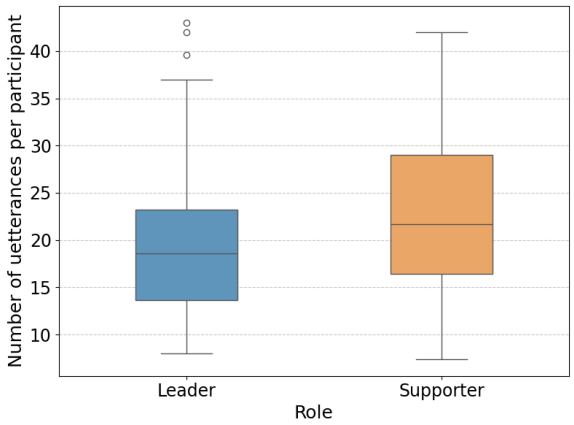

Figure 6: Distribution of the number of utterances per worker. The right figure shows the number of utterances for leaders, and the left figure shows it for supporters. If a worker participated in multiple experiments, the average number of utterances is calculated.

The *suggest* act represents the proposal of ideas or directions, meaning that a higher number of *suggest* acts implies more idea proposals. For a more detailed analysis, we performed a Brunner-Munzel test between the low-openness and high-openness groups for the count of each dialogue act. To address the issue of multiple comparisons, we applied the Bonferroni correction. As a result, there was a statistically significant difference ($p < 0.01$) in the number of *suggest* acts between the low-openness and high-openness groups.

The analysis revealed that individuals with low openness make fewer suggestions, such as ideas, during dialogue. From this result, it is considered necessary to create an environment that facilitates idea generation when supporting the creativity of people with low openness. For example, as future work, deeper insights could be gained by conducting experiments and analyses focusing on what kind of dialogue facilitates idea generation for partners with low openness.

4.2. Characteristics of the Number of Utterances per Worker

We analyze the differences in the number of utterances uttered among workers. Figure 6 shows the distribution of the number of utterances per worker, separated by role. The vertical axis represents the number of utterances a worker uttered in a single dialogue. If a worker participated in multiple experiments, we calculated the average number of utterances. This suggests that supporters generally utter a higher number of utterances than leaders. Furthermore, for leaders, the number of utterances varied greatly among workers, with the lowest being 8.0 and the highest being 43.0. Similarly, for supporters, the number of utterances also varied significantly, ranging from 7.4 for the least talkative worker to 42.0 for the most talkative. In other words, we found that there are talkative and less talkative individuals among the workers.

The analysis revealed that there are talkative and less talkative individuals among the workers, suggesting the need for personalizing dialogue systems. For example, strategies could include the dialogue system not only replying when the leader inputs text, but also actively initiating conversation and leading the dialogue.

4.3. Cluster Analysis of the Leader's Writing Workflow

To investigate the characteristics of the leader's writing behavior in each dialogue, we performed cluster analysis. For the analysis, we first filtered the edit actions of each session to extract only those records that occurred within the 30-minute time frame. Note that data where the dialogue start time was unknown, data where the story was written on the interface before the dialogue started, or data where a second story was written were excluded from this analysis.

Next, we divided this 30-minute data into 10-second intervals and focused on the logs in each interval. Then, for each 10-second interval, we defined the position where the leader was writing as a relative progress P_i against the entire story, using the following equation:

$$P_i = \frac{S_i}{S_{\text{total}}} \quad (1)$$

where S_i is the sentence number being written in interval i, and S_{total} is the total number of sentences in the completed story. For example, if the 3rd sentence was written in a specific interval i, and the

completed story in that dialogue has 10 sentences, P_i would be 0.3. It should be noted that if there is no writing log in an interval i, $P_i = P_{i-1}$. If there is no writing log in the first interval, $P_i = 0$.

To cluster these time-series data, we used the TimeSeriesKMeans from the tslearn library (Tavenard et al., 2020) in Python. The optimal number of clusters was determined to be 6 based on the elbow method, considering a range of 2 to 10 clusters. Figure 7 shows the clusters. The black line indicates the center of each cluster. The proportion of time-series data assigned to each cluster was 12.9% for Cluster 0, 46.0% for Cluster 1, 7.7% for Cluster 2, 23.3% for Cluster 3, 5.2% for Cluster 4, and 5.0% for Cluster 5.

In Cluster 0, the P value drops significantly during the middle to late stages. This suggests that workers revise the beginning of the story during the middle to late stages of the dialogue. In Cluster 1, this likely represents a style where the story is written sequentially from the first sentence right after the dialogue begins. In Cluster 2, the P value decreases before the end of the dialogue. This suggests that workers review and revise the first half of the story at the end. In Cluster 3, similar to Cluster 2, this suggests workers are reviewing their work, as they are writing in the first half of the story during the final phase of the dialogue. In Cluster 4, the last sentence of the story is written near the beginning of the dialogue. This suggests a writing style that starts from the story's conclusion. In Cluster 5, almost no story writing occurs in the first half. This suggests a style where workers first discuss thoroughly before starting to write the story.

The results showed that there is diversity in how co-creation progresses. Co-creation systems might be able to promote more efficient creative progress by personalizing the creative process. However, determining what kind of progress is most suitable needs to be clarified in future research. For instance, future studies could conduct comparative experiments using AI agents designed to prompt different creative strategies: one that encourages a sequential approach, as seen in Cluster 1, and another that prompts users to think backward from the ending, as seen in Cluster 4. Investigating how these distinct AI behaviors impact users will provide crucial insights for realizing personalized human–AI co-creative dialogue systems.

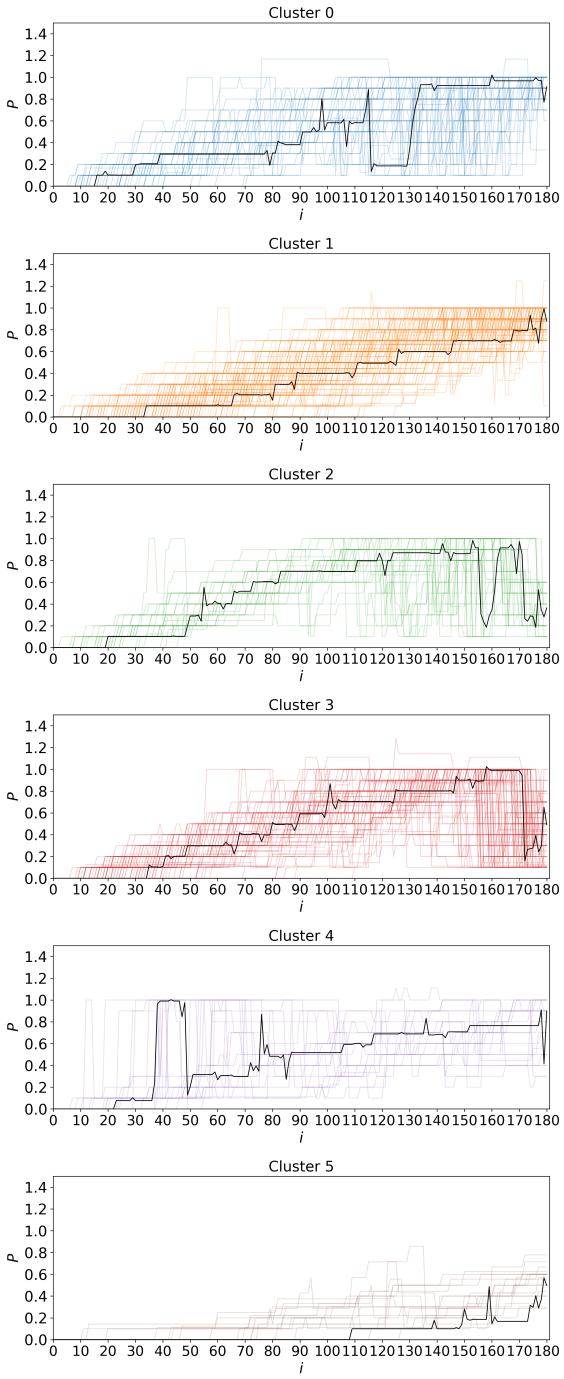

Figure 7: Cluster analysis of writing behavior. The horizontal axis represents the interval (a value close to 0 indicates the beginning of the dialogue, and a value close to 180 indicates the time near the end), and the vertical axis represents the writing position in the story at that interval. The black line represents the center of each cluster.

5. Conclusion

In this study, we collected the human–human Story Co-Creation Dialogue (StoryCCDial) dataset to realize a personalized co-creation dialogue system. We annotated the collected dialogues with dialogue act tags. In our analysis, we focused on individual characteristics (personality traits and number of utterances). The results showed that individuals with low openness made fewer suggestions during the dialogue compared to those with high openness. We also found that the number of utterances var-

ied greatly among individuals. Furthermore, we analyzed the edit histories. This analysis revealed that there is diversity in the co-creation workflow. In future work, we aim to achieve a personalized dialogue system by gaining deeper insights, such as the relationship between the postsurvey results and individual characteristics.

6. Limitation

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. Our approach is positioned within the traditional paradigm that treats human communication as the gold standard, aligning with established datasets such as MultiWOZ (Budzianowski et al., 2018) and ReDial (Li et al., 2018). We acknowledge the perspective that human–human communication may fundamentally differ from human–AI interactions; however, exploring these differences remains a subject for future investigation. 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, our analysis of the edit and dialogue histories focused exclusively on user behaviors and utterances during the co-creation process. However, because this dataset also includes the final generated stories and subjective evaluations of the co-creation experience from postsurvey, future work could investigate the relationship between the interactions during co-creation and task outcome.

7. Ethical Considerations

The workers for the co-creation dialogue were recruited via crowdsourcing and paid an appropriate wage. Furthermore, the dataset will be processed to ensure individuals cannot be identified before being released.

8. Bibliographical References

Khalid Alnajjar and Hannu Toivonen. 2021. Computational generation of slogans. *Natural Language Engineering*, 27(5):575–607.

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.

Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. 2015. User personality and the new user problem in a context-aware point of interest recommender system. In *Information and Communication Technologies in Tourism 2015: Proceedings of the International Conference in Lugano, Switzerland, February 3-6, 2015*, pages 537–549. Springer.

Gian Vittorio Caprara, Shalom Schwartz, Cristina Capanna, Michele Vecchione, and Claudio Barbaranelli. 2006. Personality and politics: Values, traits, and political choice. *Political psychology*, 27(1):1–28.

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.

Sahraoui Dhelim, Nyothiri Aung, and Huansheng Ning. 2020. Mining user interest based on personality-aware hybrid filtering in social networks. *Knowledge-Based Systems*, 206:106227.

Mehdi Elahi, Matthias Braunhofer, Francesco Ricci, and Marko Tkalcic. 2013. Personality-based active learning for collaborative filtering recommender systems. In *AI* IA 2013: Advances in Artificial Intelligence: XIIIth International Conference of the Italian Association for Artificial Intelligence, Turin, Italy, December 4-6, 2013. Proceedings 13*, pages 360–371. Springer.

Natsumi Ezure and Michimasa Inaba. 2025. The relationship between dialogue acts and idea generation in human–human collaborative story writing. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 447–460, Hanoi, Vietnam. Association for Computational Linguistics.

Somayeh Fatahi, Mina Mousavifar, and Julita Vassileva. 2023. Investigating the effectiveness of persuasive justification messages in fair music recommender systems for users with different personality traits. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 66–77.

Gregory J Feist. 1998. A meta-analysis of personality in scientific and artistic creativity. *Personality and social psychology review*, 2(4):290–309.

Lorenzo Gatti, Gözde Özbal, Oliviero Stock, and Carlo Strapparava. 2017. Automatic generation of lyrics parodies. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 485–491.

Francesco Gelli, Xiangnan He, Tao Chen, and Tat-Seng Chua. 2017. How personality affects our likes: Towards a better understanding of actionable images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1828–1837.

Komei Hiruta, Ryusuke Saito, Taro Hatakeyama, Atsushi Hashimoto, and Satoshi Kurihara. 2022. Conditional gan for small datasets. In *2022 IEEE International Symposium on Multimedia (ISM)*, pages 278–281. IEEE.

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

Rong Hu and Pearl Pu. 2010. A study on user perception of personality-based recommender systems. In *User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010. Proceedings 18*, pages 291–302. Springer.

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)*.

Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. 2018. Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers*, 20:1241–1265.

Jongeun Kim, MinChung Kim, and Taehwan Kim. 2023. Effective slogan generation with noise perturbation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3998–4002.

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.

Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, Hongyan Jiang, and Allen Tan. 2019. Personality driven multi-task learning for image aesthetic assessment. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 430–435. IEEE.

Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. 2020. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing*, 29:3898–3910.

Ruilun Liu and Xiao Hu. 2020. A multimodal music recommendation system with listeners' personality and physiological signals. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 357–360, New York, NY, USA. Association for Computing Machinery.

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.

Robert R McCrae. 1987. Creativity, divergent thinking, and openness to experience. *Journal of personality and social psychology*, 52(6):1258.

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.

Vitobha Munigala, Abhijit Mishra, Srikanth G Tamilselvam, Shreya Khare, Riddhiman Dasgupta, and Anush Sankaran. 2018. Persuaide! an adaptive persuasive text generation system for fashion domain. In *Companion Proceedings of the The Web Conference 2018*, pages 335–342.

Orestis Nalmpantis and Christos Tjortjis. 2017. The 50/50 recommender: a method incorporating personality into movie recommender systems. In *Engineering Applications of Neural Networks: 18th International Conference, EANN 2017, Athens, Greece, August 25–27, 2017, Proceedings*, pages 498–507. Springer.

Nafis Neehal and M. A. Mottalib. 2019. Prediction of preferred personality for friend recommendation in social networks using artificial neural network. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6.

Tien T Nguyen, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2018. User personality and user satisfaction with recommender systems. *Information Systems Frontiers*, 20:1173–1189.

Huansheng Ning, Sahraoui Dhelim, and Nyothiri Aung. 2019. Personet: Friend recommendation system based on big-five personality traits and hybrid filtering. *IEEE Transactions on Computational Social Systems*, 6(3):394–402.

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.

Atsushi Oshio, ABE Shingo, and Pino Cutrone. 2012. Development, reliability, and validity of the japanese version of ten item personality inventory (tipi-j). *Japanese Journal of Personality*, 21(1).

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.

Evelien Perik, Boris de Ruyter, Panos Markopoulos, and Berry Eggen. 2004. The sensitivities of user profile information in music recommender systems. *Proceedings of private, security, trust*, 137.

Peter J Rentfrow and Samuel D Gosling. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236.

Jesus F Salgado. 2002. The big five personality dimensions and counterproductive behaviors. *International journal of selection and assessment*, 10(1-2):117–125.

Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: 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.

Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. 2020. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6.

Marko Tkalcic, Matevz Kunaver, Jurij Tasic, and Andrej Košir. 2009. Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*, pages 30–37.

Hsin-Chang Yang and Zi-Rui Huang. 2019. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems*, 165:157–168.

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.

Virgil Zeigler-Hill and Sean Monica. 2015. The hexaco model of personality and video game preferences. *Entertainment Computing*, 11:21–26.

Xulin Zhou, Takuma Ichikawa, and Ryuichiro Higashinaka. 2024. Collecting and analyzing dialogues in a tagline co-writing task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3507–3517, Torino, Italia. ELRA and ICCL.

9. Language Resource References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. 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).

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.

Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9748–9758, Red Hook, NY, USA. Curran Associates Inc.

A. Prompt for annotation

Initial dialogue act annotation was performed using GPT-4-turbo. The prompt for annotation consisted of three parts (Figure 8): 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.

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