Analyzing Differences in Subjective Annotations by Participants and Third-party Annotators in Multimodal Dialogue Corpus

Kazunori Komatani **Ryu Takeda**
SANKEN, Osaka University
Ibaraki, Osaka 567-0047, Japan
{komatani,rtakeda}@sanken.osaka-u.ac.jp

Shogo Okada
JAIST
Nomi, Ishikawa 923-1292, Japan
okada-s@jaist.ac.jp

Abstract

Estimating the subjective impressions of human users during a dialogue is necessary when constructing a dialogue system that can respond adaptively to their emotional states. However, such subjective impressions (e.g., how much the user enjoys the dialogue) are inherently ambiguous, and the annotation results provided by multiple annotators do not always agree because they depend on the subjectivity of the annotators. In this paper, we analyzed the annotation results using 13,226 exchanges from 155 participants in a multimodal dialogue corpus called Hazumi that we had constructed, where each exchange was annotated by five third-party annotators. We investigated the agreement between the subjective annotations given by the third-party annotators and the participants themselves, on both per-exchange annotations (i.e., participant's sentiments) and per-dialogue (-participant) annotations (i.e., questionnaires on rapport and personality traits). We also investigated the conditions under which the annotation results are reliable. Our findings demonstrate that the dispersion of third-party sentiment annotations correlates with agreeableness of the participants, one of the Big Five personality traits.

1 Introduction

To achieve adaptive human-machine (or human-robot) dialogue, it is necessary to estimate the human user's subjective impressions and emotions during the dialogue. The user's satisfaction with the dialogue can be increased by appropriately changing the dialogue content in accordance with the user's emotions. Estimated subjective impressions and emotions can also be utilized to evaluate the dialogue.

The difficulty here is that such impressions and feelings are inherently subjective, and it is impossible to objectively determine unique references for subjective content. References are necessary for training and evaluating machine learning models. Even when manual annotations are performed, the results among annotators do not always agree, which is a common problem in annotations of subjective labels.

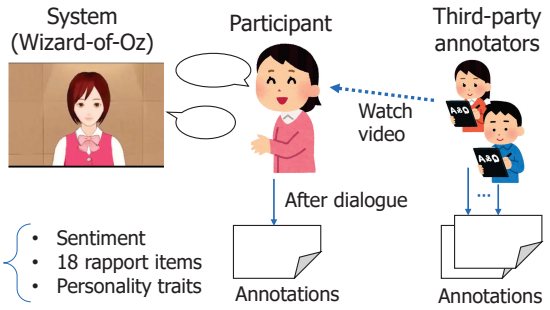

Figure 1: Subjective annotations given by participants themselves and by third-party annotators.

In this paper, we analyze the disagreement among human annotation results, especially the differences between annotations given by participants themselves and by third-party annotators (Fig. 1). Specifically, we conducted investigations on per-exchange annotations, i.e., sentiments, and per-dialogue (per-participant) annotations, i.e., questionnaires measuring the participant's rapport and personality traits. We used the Osaka University Multimodal Dialogue Corpus Hazumi, which we had previously constructed (Komatani and Okada, 2021), for the analysis. Our findings show that third-party annotators tend to give subjective annotations on the basis of their rather simple impressions compared to the participants themselves, who may not always fully express their inner states during the dialogue. We also clarify why automatic estimation performances of per-exchange sentiments from multimodal features differ between cases in which the reference sentiments were given by the participants themselves and by the third-party annotators, where the latter usually obtains better performances.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 104–113
September 11–15, 2023.

We then investigate the conditions under which estimated sentiments given by third-party annotators would be reliable by examining the dispersion of the annotation results. The estimation of users' sentiments based on multimodal data with machine learning will never be perfect, so it would be helpful to know whether the estimation results can be reliable for each user on the basis of other information sources. In this paper, after showing how the dispersion of sentiments given by third-party annotators correlates with machine learning performance, we demonstrate that this dispersion is negatively correlated with one of the personality traits, namely, agreeableness. This finding indicates that a personality trait can be a useful clue for determining the reliability of the sentiment estimation results.

2 Related Work

We here describe related studies on adaptive dialogue systems, emotion recognition, datasets of multimodal dialogues, reference labels for subjective annotations, and personality traits, in that order.

It is essential that dialogue system responses be adaptive to user states. In task-oriented dialogues, task success rates can be improved and the number of turns to task completion can be reduced by adapting system responses in accordance with several user types (Komatani et al., 2005). As for non-task-oriented dialogues, personalization based on the user's domain expertise has been attempted (Ferrod et al., 2021). System responses are preferably based on various modalities such as vision and prosody in addition to textual input. A variety of studies have examined text-based chatbots based on large pre-trained language models (e.g., (Adiwardana et al., 2020; Roller et al., 2021)). Currently, studies on dialogue systems have been actively expanded from the text-based perspective to a multimodal one, as evidenced by a recent dialogue competition using a humanoid robot (Minato et al., 2022).

User impressions (such as emotions) can be an important clue for adaptive dialogue systems. In particular, adapting to the user's emotions is essential for social interaction (Barros et al., 2021). Moreover, different types of information, including multimodal information (e.g., vision and prosody), can be utilized to recognize the user's emotions, as can physiological signals (Katada et al., 2022; Wu et al., 2022). In this paper, emotion is treated as sentiments per exchange.

A famous multimodal dialogue corpus with emotion labels is the IEMOCAP dataset (Busso et al., 2008), which contains dialogues between actors in role-playing scenarios. The Emotional Dyadic Motion CAPture (IEMOCAP) dataset is a well-known dataset used to recognize emotion during dialogues (Busso et al., 2008). It is a well-controlled dataset in the sense that data were collected by asking actors to speak with designated emotions. Therefore, this dataset contains objective reference labels for each emotion, i.e., the designated emotions. In contrast, our Hazumi dataset (Komatani and Okada, 2021) utilized in this paper consists of natural and spontaneous dialogues. Thus, there are no objective reference labels. We opted to use this dataset because our objective is to analyze the differences between several manual annotation results and discuss reference labels for subjective annotations.

Prior studies in the fields of social signal processing and affective computing have examined how to determine the ground truth of subjectively assigned labels (Spodenkiewicz et al., 2018; Bourvis et al., 2021; Maman et al., 2022). Maman et al. (2022) proposed three strategies for utilizing self-assessment labels and external assessment labels in training data for two dimensions of a group engagement state (called cohesion) and compared their prediction performances. Wang et al. (2023) recently proposed a method to train a classifier that fits better with the annotation results in medical binary classification tasks. In this paper, we do not train a classifier but analyze what happened in a multimodal dialogue data. We also extend analysis from single to several subjective annotations, i.e., per-exchange annotation and per-dialogue annotations.

Emotion depends on individual users, e.g., their personality traits (such as the Big Five (Goldberg, 1990)). Personality traits also play an important role in a variety of user-adapted interactions (Mairesse and Walker, 2010; Mota et al., 2018; Fernau et al., 2022; Yamamoto et al., 2023). The personality traits of a robot and human interlocutors are known to be effective for engagement estimation in human-robot interactions (Salam et al., 2017), and correlation between the engagement and the personality traits given per dialogue has been investigated in human-robot and human-human interactions (Celiktutan et al., 2019). In this work, we

Table 1: Hazumi versions and corresponding annotations.

Version	Recording environment	No. of participants (dialogues)	No. of exchanges	Self-sentiment	Third-party sentiment	18 rapport items	Personality traits
Hazumi1712	in-person	29	2,422		√		
Hazumi1902		30	2,514	√	√	√	
Hazumi1911		30	2,859	√	√	√	√
Hazumi2010	online	33	2,798		√	√	√
Hazumi2012		63	5,334		√	√	√
Hazumi2105		29	2,235		√	√	√
Total		214	18,162				

comprehensively analyzed the relationship between the user's personality traits on the basis of per-dialogue questionnaire results and per-exchange sentiments.

3 Target Corpus

We utilized the multimodal dialogue corpus Hazumi, which we had previously constructed (Komatani and Okada, 2021). It is a dataset that can be used extensively for research and development purposes[1]. Table 1 lists the various versions of the Hazumi corpus along with their recording environments, numbers of participants and exchanges, and annotations. It has six versions: 1712, 1902, 1911, 2010, 2012, and 2105, where the numbers correspond to the year and month the data collection started; for example, the collection of Hazumi1911 data began in November 2019. The first three versions were collected in-person and the following three were collected online due to the COVID-19 pandemic. Each dialogue lasted approximately 15 to 20 minutes.

The annotation unit at the utterance level is the exchange. An exchange is defined from the beginning of a system utterance to the beginning of the next system utterance. The data contain 18,162 exchanges in total; the mean duration was 13.10 seconds and its standard deviation was 7.80.

3.1 Dialogue data details

In Hazumi, the system used by the participants for talking was MMDAgent (Lee et al., 2013), which was operated by the Wizard-of-Oz (WoZ) method in which the virtual agent was controlled by a human operator (Wizard) located in another room. The Wizard controlled a graphical user interface built for this task while remotely observing the participants. Since the operators were trained to select the next utterance while the participant was still speaking (approximately ten seconds), there was a short wait time before the agent started responding.

The dialogue was chit-chat, meaning there was no specific task to be completed. The conversations were in Japanese and spanned several topics such as travel and movies. The Wizard attempted to select utterances that would engage the participants for a longer time. Specifically, the Wizard changed topics when the participants seemed uninterested, and listened when the participants seemed interested and were actively talking.

The participants were recruited from the general public through a recruiting agency for the in-person collection and through crowdsourcing for the online collection. A total of 214 participants (99 men, 115 women) were included, ranging in age from their 20s to 70s. They were given no special instructions, such as requests to act out their emotions strongly. Data were collected only from participants who signed a consent form that stated the data could be distributed to researchers for research and development purposes.

3.2 Subjective annotations

Manual annotations were given at the utterance and dialogue levels. The right half of Table 1 shows the types of subjective annotations and the Hazumi versions to which they were annotated.

3.2.1 Per-exchange annotations

Sentiment is scored on a 7-point scale representing how much the participant enjoyed the dialogue. Annotators gave it once per exchange, while watching the recorded videos of the dialogues. The sentiment annotation given by the third-party annotators is called *third-party sentiment*. For Hazumi1902 and Hazumi1911, the sentiment was also given by the participants themselves, which is called *self-sentiment*. They watched the recorded video and provided annotations immediately after their dialogue.

[1]The corpus has been distributed by the Informatics Research Data Repository at the National Institute of Informatics (NII-IDR). https://www.nii.ac.jp/dsc/idr/en/rdata/Hazumi/

Table 2: Cronbach's alpha values among five third-party annotators for per-dialogue annotations.

	Personality traits (Big Five)					Average of 18 rapport items
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	
Hazumi1911	0.835	0.761	0.622	0.620	0.696	0.876
Hazumi2010	0.911	0.791	0.560	0.697	0.883	0.883
Hazumi2012	0.867	0.827	0.598	0.599	0.747	0.856
Hazumi2105	0.903	0.843	0.663	0.645	0.786	0.813

3.2.2 Per-dialogue annotations

As per-dialogue annotations, participants answered two questionnaires after completing their dialogue: *18 rapport items*, which measured their rapport in the dialogue, and *Personality traits*, which examined their personality traits. Five third-party annotators also answered the same questionnaires about the participants from a third-party perspective after watching the recorded videos of the dialogues (i.e., they did not just read the transcribed texts).

The *18 rapport items* questionnaire was developed by social psychologists and originally consisted of 18 English adjectives[2] (Bernieri et al., 1996). It aims to examine the interlocutor's rapport and the results indicate how the dialogue was perceived. We utilized 18 questionnaire items with the 18 adjectives translated and converted into Japanese sentences (Kimura et al., 2005), such as "1. The dialogue was well-coordinated," "2. The dialogue was boring," and "18. The dialogue was slow." Each item is scored on an 8-point scale.

The second questionnaire asked about the participants' *personality traits* modeled on the Big Five, that is, extraversion, agreeableness, conscientiousness, neuroticism, and openness (Goldberg, 1990; Vinciarelli and Mohammadi, 2014). We used the 10-item personality inventory translated into Japanese (TIPI-J) (Oshio et al., 2012), which measures the Big Five with ten items. The items are scored on a 7-point scale, with two questions for each of the traits, one of which is an inverted item. Each of the Big Five scores is the sum of the two question items, one of which corresponds to the inverted item subtracted from 8 (i.e., the minimum is 2 and the maximum is 14).

As a preliminary analysis, Table 2 shows the Cronbach's alpha values among the five third-party annotators for the two kinds of per-dialogue annotations. An annotation result is considered consistent if the Cronbach's alpha is greater than 0.8. As we can see, extraversion, agreeableness, and openness tended to be around 0.8 or above, while conscientiousness and neuroticism tended to be below 0.8. This is consistent with the results of personality trait annotation agreement rates in other studies (Aran and Gatica-Perez, 2013). The values of the Cronbach's alpha for the average of the 18 rapport items also tended to be consistent.

[2]All adjectives appear in Table 4.

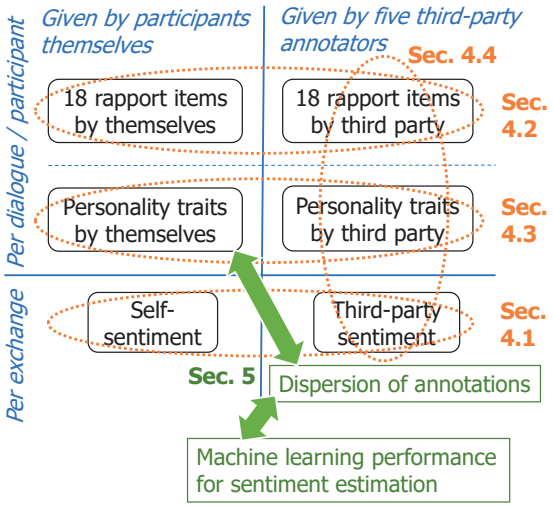

Figure 2: Positioning of the analyses.

4 Analyses on Relationship Between Annotations Given by Participants and Third-Party Annotators

We analyzed the correlations between the manual annotations given by the participants themselves and by five third-party annotators. Sentiments were analyzed using Hazumi1902 and Hazumi1911 due to their availability (see Table 1). As for the two annotations given per dialogue, we used the data of the four versions after Hazumi1911, which consist of 13,226 exchanges from 155 participants. Figure 2 depicts the positioning of the analyses we conducted.

If any correlation is found between two metrics corresponding to the annotations, it will provide useful insights for the machine learning design. For example, it would be effective to use one of the metrics as input when estimating the other by machine learning. The correlation would also be

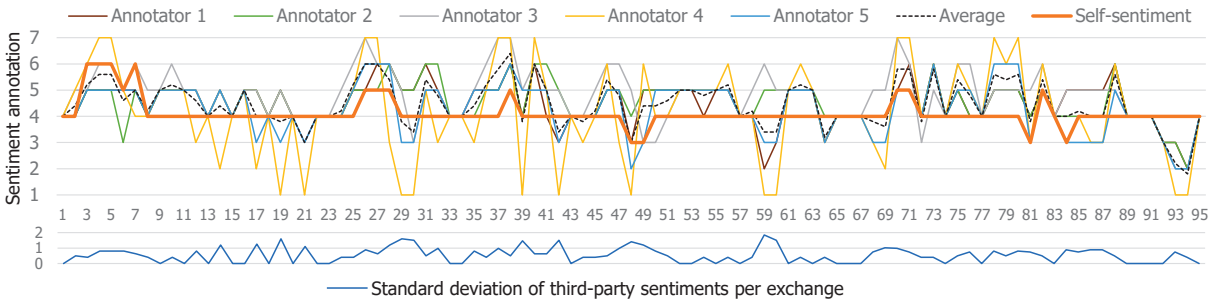

Figure 3: Example of sentiment annotation results and standard deviations (participant ID: 1911M4001).

helpful in designing multi-task learning with deep neural networks in which some layers are shared (Hirano et al., 2019). The two metrics can be utilized together to improve the machine learning performance.

4.1 Sentiment

Figure 3 shows an example of the sentiment annotation results for a male participant in his 40s (participant ID: 1911M4001). Horizontal and vertical axes indicate time in units of exchange and the annotation results on the 7-point scale, respectively. The solid lines in different colors represent the third-party sentiments by the five annotators (1 male, 4 female; Annotator 5 was male). The thick orange line in the center is self-sentiment, which does not agree with the third-party sentiments. The third-party sentiments by the five annotators share certain trends but do not completely agree. The correlation coefficient between the self-sentiment and the average of the third-party sentiments was 0.45. The figure also shows standard deviations of the third-party sentiments per exchange at the bottom, which will be used in Section 5.

Table 3 shows the correlation coefficients between self-sentiments and third-party sentiments, which were calculated per participant. The macro average of all correlation coefficients was 0.43. The maximum was 0.79 and the minimum was 0.01, indicating large individual differences. These results clarify that the self-sentiments and third-party sentiments are not necessarily correlated, as reported in (Truong et al., 2012).

This is why automatic estimation performances from multimodal features differ between cases in which self-sentiments and third-party sentiments are used as the references (Katada et al., 2022), where the latter obtained better performances. Third-party sentiments can be perceived from outside the participants, which suggests that computers attempting to estimate the sentiments can utilize the same information that the third-party annotators use. Self-sentiment is more difficult to estimate because it is not necessarily perceivable from the outside, even by human third-party annotators. Additional use of physiological signals has thus improved the estimation performance of self-sentiment (Katada et al., 2022). The signals can be regarded as extra information that third-party annotators can perceive.

Table 3: Correlation between self-sentiments and third-party sentiments.

	No. of participants	Macro average	(max., min.)
Hazumi1902	30	0.45	(0.69, 0.11)
Hazumi1911	30	0.41	(0.79, 0.01)
Total	60	0.43	(0.79, 0.01)

We also confirmed here that the correlation coefficients differ among participants and that the sentiment annotations results differ among the third-party annotators. We therefore attempted to use the deviation of the third-party sentiments in Section 5.

4.2 18 rapport items

We investigated the correlation between the answers by participants themselves and the averages of third-party annotators for each of the 18 rapport items. Table 4 lists the correlation coefficients in descending order. Excluding the three below the solid line, all correlations were statistically significant ($p < 0.05$). The correlation between the averages of the correlation coefficients was 0.34 (bottom line), and it was also statistically significant ($p = 0.023$).

Thus, the averaged answers to the 18 rapport items, which correspond to the posterior evaluation of the dialogue, showed a correlation between the participants themselves and the averages of the third-party annotations. The results in Table 4

Table 4: Correlation coefficients of all 18 rapport items between self- and third-party annotations.

5*	unsatisfying	0.38
9	engrossing	0.35
2*	boring	0.32
17	worthwhile	0.29
8*	awkward	0.27
16*	dull	0.25
10*	unfocused	0.23
6*	uncomfortably paced	0.23
1	well-coordinated	0.22
12*	intense	0.21
11	involving	0.21
14	active	0.20
4	harmonious	0.20
7*	cold	0.19
18*	slow	0.17
13	friendly	0.13
15	positive	0.09
3	cooperative	0.07
Average of 18 items		0.34

'*' denotes inverted items.

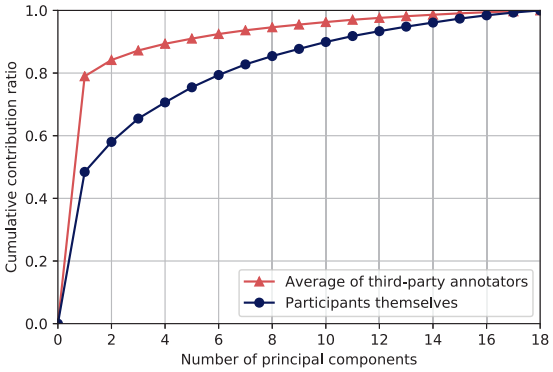

Figure 4: Cumulative contribution ratios by PCA for 18 rapport items.

also suggest that the upper-level items are mostly related to the content of the conversation (e.g., unsatisfying, engrossing, and boring). In contrast, the lower-level items are related to the feeling and atmosphere of the dialogue (e.g., friendly, positive, and cooperative).

We also applied principal component analysis (PCA) to the results of the answers to the 18 rapport items for each of those by participants themselves and the averages by the third-party annotators. Table 4 lists the cumulative contribution ratio of the PCA. The contribution ratios of the first principal components were 0.790 and 0.484 for the answers by the third-party annotators and participants themselves, respectively. These results indicate that one dimension could explain about 80% of the answers by the third-party annotators; in other words, the third-party annotators tended to answer the 18 items on the basis of rather simple impressions of positive or negative. In contrast, the participants presumably answered after considering more complicated inner impressions of the dialogue that they were actually participating in.

4.3 Personality traits

Table 5 shows the correlations between the personality traits reported by the participants themselves and the averages given by the five third-party annotators. The correlation coefficients for extraversion were consistently large and statistically significant among the versions, but the overall tendency appears to be that the other personality traits by the participants themselves do not necessarily correlate with the averages by the third-party annotators. Openness had the next largest correlation coefficient, followed by conscientiousness. The reason extraversion had high correlation coefficients is that it (by definition) tends to be more easily expressed during dialogue. This result is consistent with an experiment in the psychology field (Borkenau et al., 2009) in which extraversion was reported to be highly consistent between self-rating and rating by others.

It makes sense that the annotation results do not necessarily correlate if the personality traits of the participants are not sufficiently expressed in the dialogue, e.g., for neuroticism and agreeableness. This is because third-party annotators do not know the participants and score personality traits based only on their impression during the dialogue.

4.4 Relation among annotation results by third-party annotators

We investigated the correlations among the above annotation results given by the third-party annotators for sentiments, 18 rapport items, and personality traits. Table 6 lists the correlation of each of the five personality traits with the averages of the 18 rapport items and sentiments. As we can see, the average of the 18 rapport items correlated with all of the five personality traits with statistical significance, especially for agreeableness, extraversion, and openness, whose correlation coefficients were 0.68, 0.53, and 0.52, respectively. Similarly, the average of sentiments correlated with three personality traits (openness, agreeableness, and extraversion) with statistical significance; their correlation coefficients were 0.36, 0.30, and 0.21, respectively. In addition, the average of the 18 rapport items

Table 5: Correlation between personality traits given by participants themselves and averages given by third-party annotators.

	No. of participants	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Hazumi1911	30	0.53	0.08	0.43	0.25	0.29
Hazumi2010	33	0.58	−0.44	0.17	0.10	0.34
Hazumi2012	63	0.39	0.19	0.11	0.14	0.19
Hazumi2105	29	0.57	0.37	0.06	0.21	0.17
Total	155	0.49	0.06	0.16	0.15	0.21

Underlined values indicate statistical significance ($p < 0.05$).

Table 6: Correlation of personality traits with 18 rapport items and sentiments. All are averages given by five third-party annotators.

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Average of 18 rapport items	0.53	0.68	0.21	−0.22	0.52
Average of sentiments	0.21	0.30	0.12	0.00	0.36

Underlined values indicate statistical significance ($p < 0.05$).

also correlated with the average of sentiments with statistical significance; its correlation coefficient was 0.55.

These results confirm that there were correlations between the three annotation results given by the five third-party annotators, thereby demonstrating that these three metrics can help each other in their estimation using machine learning. For example, in a dialogue where the participant seemed to enjoy talking, the average of the sentiments was high, the average of the 18 rapport items was also high, and the participant's extraversion, cooperativeness, and openness also seemed high. This simple tendency is echoed our discussion about the results of the PCA analysis in Section 4.2: that is, the third-party annotators tended to annotate on the basis of rather simple impressions of positive or negative.

5 Analyses on Dispersion of Third-Party Sentiments

We here focus on the dispersion of sentiments given by the five third-party annotators (third-party sentiments). We discuss the conditions under which the third-party sentiments would be reliable.

5.1 Formulating dispersion of third-party sentiments

The bottom line in Fig. 3 shows the standard deviations of the third-party sentiments for each exchange. Using this as a basis, we formulate the dispersion of third-party sentiments as the averages of the standard deviations, as follows.

Let $dispersion(i)$ denote the dispersion of third-party sentiments for a participant i (i.e., dialogue). Values a_{ijk} denote third-party sentiments for the j-th exchange ($j = 1, ..., J_i$) in the dialogue with participant i by the k-th third-party annotator ($k = 1, ..., K$). The values of sentiments are annotated on a 7-point scale, i.e., $a_{ijk} \in \{1, ..., 7\}$. J_i denotes the total number of exchanges in the dialogue with participant i, which is 95 in the example in Fig. 3. K denotes the number of third-party annotators, i.e., $K = 5$. Standard deviations of the annotated sentiments

$$stdev(i, j) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(a_{ijk} - \overline{a_{ij}})^2} \quad (1)$$

can be calculated per exchange. Here, $\overline{a_{ij}}$ denotes the averages of the third-party sentiments given by K annotators for the j-th exchange. We define $dispersion(i)$ of third-party sentiments for a participant i (i.e., dialogue) as the average of the standard deviations $stdev(i, j)$, i.e.,

$$dispersion(i) = \frac{1}{J_i}\sum_{j=1}^{J_i} stdev(i, j). \quad (2)$$

5.2 Relationship between dispersion and machine learning performance

Here, we discuss the relationship between the dispersion of third-party sentiments and the performance of machine learning. This explains why we focused on the dispersion.

It is known empirically that machine learning performs better when the manual annotations agree more. For example, in an emotion recognition task for spoken utterances, it was reported that the recognition performance based on machine learning was

Table 7: Correlation between the dispersion of third-party sentiments and personality traits.

	No. of participants	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Hazumi1911	30	0.24	−0.44	0.16	0.12	0.27
Hazumi2010	33	0.38	−0.38	−0.15	0.11	0.04
Hazumi2012	63	−0.13	−0.20	0.00	0.08	−0.05
Hazumi2105	29	−0.20	−0.13	0.29	−0.04	0.03
Total	155	−0.05	−0.26	−0.05	0.03	−0.04

Underlined values indicate statistical significance ($p < 0.05$).

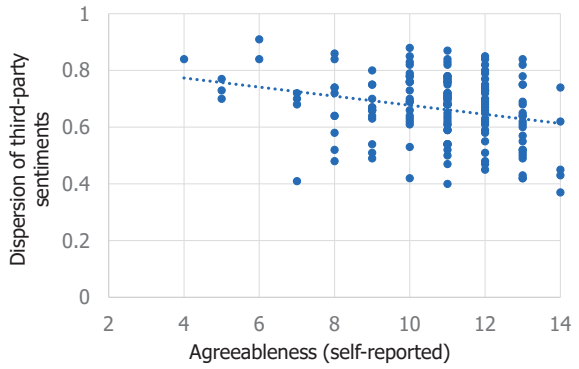

Figure 5: Correlation between the dispersions and the scores of agreeableness for each participant.

better on units to which multiple annotators gave the same labels (Seppi et al., 2008). As a preliminary investigation, we also calculated the mean square error of the sentiment estimation results as the regression from multimodal features (Katada et al., 2020) using 2,468 exchanges in Hazumi1911, where the references were the average of third-party sentiments. The correlation coefficient between the mean square errors and the standard deviations of the third-party sentiments per exchange was 0.342, which is statistically significant ($p = 1.57 \times 10^{-68}$). In other words, the error in machine learning results tends to be larger for exchanges with large standard deviations of third-party sentiments.

These results suggest that the sentiment estimation performance based on machine learning tends to be lower for parts with large deviations in human judgment.

5.3 Correlation between dispersion and personality traits

We investigated the correlation between the dispersion of third-party sentiments and the personality traits of each participant. The personality traits utilized here are those reported by the participants themselves. Table 7 lists the correlation coefficients between each of the five personality traits and the dispersions of third-party sentiments per participant (calculated by Eq. (2)), for each of the four versions and in total. We can see here that agreeableness negatively correlates with the dispersion of third-party sentiments. Specifically, the correlation coefficient with agreeableness was -0.26 for the total, which is statistically significant ($p = 9.1 \times 10^{-4}$).

Figure 5 shows the dispersions of the third-party sentiments and the scores of agreeableness. Each point denotes 155 participants from Hazumi1911 to Hazumi2105. Horizontal and vertical axes denote the score of agreeableness and the dispersions of the third-party sentiments for each participant. We can see here that there is a negative correlation between these two metrics. In other words, there were fewer dispersions of the third-party sentiments for the participants who recognized themselves as more agreeable. This result can be interpreted as a phenomenon that the more agreeable the participant is, the more he/she tries to express his/her sentiments in a way that the interlocutor (and thus the third-party annotators) can recognize.

The results of the sentiment estimation for highly agreeable users thus tend to be reliable, given the low dispersion of the third-party sentiments, which tend to correlate with machine learning performance, as discussed in Section 5.2.

6 Conclusion

In this paper, we investigated the correlation of subjective annotation results between the participants themselves and five third-party annotators. We found that some are correlated, which will potentially be useful in machine learning to estimate one of the annotation targets, such as the participants' sentiments, their evaluation of dialogues (18 rapport items), or their personality traits.

We also investigated the dispersion of the sentiments given by the five third-party annotators. We showed that a difference in annotation results correlates with the estimation error of machine learning and found that the dispersion was negatively cor-

related with agreeableness, one of the Big Five personality traits.

These results can provide insights into the development of adaptive dialogue systems: specifically, a personality trait can be used as a clue to determine whether or not to rely on the sentiment recognition results. One of our future works is to estimate the user's personality traits before and during the dialogue. The system can then utilize the personality trait to decide how actively to adapt to the user on the basis of the discussion in this paper. Personality traits such as neuroticism are not expressed by users during dialogues such as chit-chat and thus are difficult for the system and third-party annotators to observe. The analyses in this paper considered all of the Big Five traits, but it will be necessary to select personality traits observable in the dialogue accordingly, e.g., extraversion.

The results presented in this paper are based on our Japanese dataset Hazumi. Various factors such as the behavior of the participants and annotators, for example, can be involved. Further investigation is needed to confirm the generalizability of the obtained results to other languages and cultures, as well as to different experimental settings including dialogue tasks and instructions to the participants.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP22H00536 and JP19H05692, and JST Moonshot R&D Grant Number JPMJPS2011.

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Oya Aran and Daniel Gatica-Perez. 2013. One of a kind: Inferring personality impressions in meetings. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 11–18.

Pablo Alves De Barros, Ana Tanevska, and Alessandra Sciutti. 2021. Affect-aware learning for social robots. In *Adjunct Proc. Conference on User Modeling, Adaptation and Personalization*, pages 130–132.

Frank J. Bernieri, John S. Gillis, Janet M. Davis, and Jon E. Grahe. 1996. Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1):110–129.

Peter Borkenau, Steffi Brecke, Christine Mottig, and Marko Paelecke. 2009. Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, 43(4):703–706.

Nadege Bourvis, Aveline Aouidad, Michel Spodenkiewicz, Giuseppe Palestra, Jonathan Aigrain, Axel Baptista, Jean-Jacques Benoliel, Mohamed Chetouani, and David Cohen. 2021. Adolescents with borderline personality disorder show a higher response to stress but a lack of self-perception: Evidence through affective computing. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 111:110095.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Na rayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. 2019. Multimodal human-human-robot interactions (MHHRI) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4):484–497.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards personality-aware chatbots. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 135–145.

Roger Ferrod, Federica Cena, Luigi Di Caro, Dario Mana, and Rossana Grazia Simeoni. 2021. Identifying users' domain expertise from dialogues. In *Adjunct Proc. Conference on User Modeling, Adaptation and Personalization*, pages 29–34.

R. Lewis Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, pages 1216–1229.

Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *Proc. International Conference on Multimodal Interaction (ICMI)*, page 85–94.

Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is she truly enjoying the conversation? Analysis of physiological signals toward adaptive dialogue systems. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 315–323.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing*.

Masanori Kimura, Masao Yogo, and Ikuo Daibo. 2005. Expressivity halo effect in the conversation about emotional episodes (in Japanese). *The Japanese Journal of Research on Emotions*, 12(1):12–23.

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

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. 2005. User modeling in spoken dialogue systems to generate flexible guidance. *User Modeling and User-Adapted Interaction*, 15(1):169–183.

Akinobu Lee, Keiichiro Oura, and Keiichi Tokuda. 2013. MMDAgent –a fully open-source toolkit for voice interaction systems–. In *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pages 8382–8385.

François Mairesse and Marilyn A. Walker. 2010. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278.

Lucien Maman, Gualtiero Volpe, and Giovanna Varni. 2022. Training computational models of group processes without groundtruth: The self- vs external assessment's dilemma. In *Companion Publication of the 2022 International Conference on Multimodal Interaction (ICMI)*, pages 18–23.

Takashi Minato, Ryuichiro Higashinaka, Kurima Sakai, Tomo Funayama, Hiromitsu Nishizaki, and Takayuki Nagai. 2022. Overview of dialogue robot competition 2022. *arXiv preprint arXiv:2210.12863*.

Pedro Mota, Maike Paetzel, Andrea Fox, Aida Amini, Siddarth Srinivasan, and James Kennedy. 2018. Expressing coherent personality with incremental acquisition of multimodal behaviors. In *Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 396–403.

Atsushi Oshio, Shingo Abe, and Pino Cutrone. 2012. Development, reliability, and validity of the Japanese version of ten item personality inventory (TIPI-J) (in Japanese). *The Japanese Journal of Personality*, 21(1):40–52.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proc. European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325, Online.

Hanan Salam, Oya Celiktutan, Isabelle Hupont, Hatice Gunes, and Mohamed Chetouani. 2017. Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, 5:705–721.

Dino Seppi, Anton Batliner, Björn Schuller, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, and Vered Aharonson. 2008. Patterns, prototypes, performance: classifying emotional user states. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 601–604.

Michel Spodenkiewicz, Jonathan Aigrain, Nadège Bourvis, Séverine Dubuisson, Mohamed Chetouani, and David Cohen. 2018. Distinguish self- and hetero-perceived stress through behavioral imaging and physiological features. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 82:107–114.

Khiet P. Truong, David A. van Leeuwen, and Franciska M.G. de Jong. 2012. Speech-based recognition of self-reported and observed emotion in a dimensional space. *Speech Communication*, 54(9):1049–1063.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Chongyang Wang, Yuan Gao, Chenyou Fan, Junjie Hu, Tin Lun Lam, Nicholas Donald Lane, and Nadia Berthouze. 2023. Learn2agree: Fitting with multiple annotators without objective ground truth. In *ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare*.

Yuyan Wu, Miguel Arevalillo Herráez, Stamos Katsigiannis, and Naeem Ramzan. 2022. On the benefits of using hidden markov models to predict emotions. In *Proc. Conference on User Modeling, Adaptation and Personalization*, pages 164–169.

Kenta Yamamoto, Koji Inoue, and Tatsuya Kawahara. 2023. Character adaptation of spoken dialogue systems based on user personalities. In *Proc. International Workshop on Spoken Dialogue System Technology (IWSDS)*.