# DSLCMM: A Multimodal Human-Machine Dialogue Corpus Built through Competitions

**Ryuichiro Higashinaka[1], Tetsuro Takahashi[2], Shinya Iizuka[1], Sota Horiuchi[1], Michimasa Inaba[3], Zhiyang Qi[3], Yuta Sasaki[4], Kotaro Funakoshi[4], Shoji Moriya[5], Shiki Sato[5], Takashi Minato[6], Kurima Sakai[7], Tomo Funayama[7], Masato Komuro[8], Hiroyuki Nishikawa[9], Ryosaku Makino[10], Hirofumi Kikuchi[10], Mayumi Usami[11]**

[1]Nagoya University, [2]Kagoshima University, [3]The University of Electro-Communications, [4]Science Tokyo, [5]Tohoku University, [6]RIKEN, [7]ATR, [8]Chiba University, [9]Meikai University, [10]Waseda University, [11]Tokyo University of Foreign Studies

**Correspondence:** higashinaka@i.nagoya-u.ac.jp

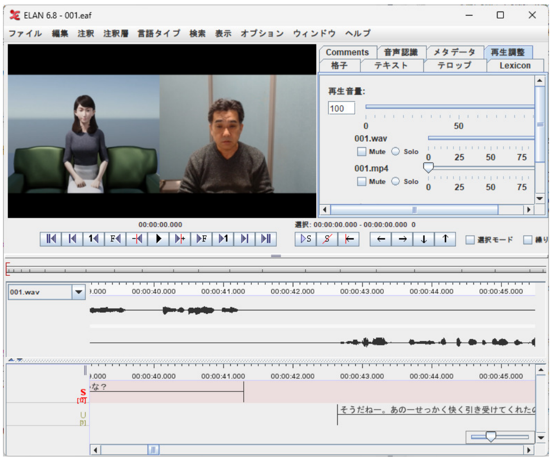
## Abstract

A corpus of dialogues between multimodal systems and humans is indispensable for the development and improvement of such systems. However, there is a shortage of human-machine multimodal dialogue datasets, which hinders the widespread deployment of these systems in society. To address this issue, we construct a Japanese multimodal human-machine dialogue corpus, DSLCMM, by collecting and organizing data from the Dialogue System Live Competitions (DSLCs). This paper details the procedure for constructing the corpus and presents our analysis of the relationship between various dialogue features and evaluation scores provided by users.

## 1 Introduction

With the advancement of large language models, the capabilities of text-based dialogue systems have improved (Shuster et al., 2022; Hudeček and Dušek, 2023; Kong et al., 2024). However, the performance of multimodal dialogue systems, which must process speech, facial expressions, and other non-verbal cues in real time, remains limited compared to human face-to-face interactions (Higashinaka et al., 2024).

To unlock the full potential of dialogue systems in society, it is essential to achieve multimodal dialogue capabilities akin to human interactions. However, there is a significant shortage of corpora to support this. While a relatively large amount of multimodal data focuses on human-to-human dialogue or human-to-Wizard of Oz (WoZ) interactions, there is a notable lack of corpora capturing dialogues between actual multimodal dialogue systems and humans. To understand how real systems and humans interact in multimodal dialogues and identify areas for improvement, a corpus of such interactions is indispensable.



な？

そうだねー。あのーせっかく快く引き受けてくれた

Figure 1: Example data in DSLCMM displayed using ELAN (Wittenburg et al., 2006).

In light of this background, we have constructed a multimodal dialogue corpus called DSLCMM. Specifically, we collected and processed data from the Dialogue System Live Competition (DSLC) series (Higashinaka et al., 2021) and organized it into a corpus. This corpus contains 1,747 dialogues between 32 multimodal dialogue systems and human users, obtained from two editions of the competition. The language of the corpus is Japanese. In addition to users' speech, the corpus contains user/system video recordings and logs of system commands for gestures and facial expressions. It also includes subjective evaluation scores from users and transcriptions of all user utterances. An example of the dataset is shown in Figure 1. Utilizing this dataset has the potential to significantly advance research on multimodal dialogue systems. The corpus will be accessible from the project pages of DSLC5[1] and DSLC6[2].

[1] https://sites.google.com/view/dslc5

[2] https://sites.google.com/view/dslc6

*Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 277–283
May 27–30, 2025. ©2025 Association for Computational Linguistics

In Section 2 of this paper, we review related work. Section 3 provides an explanation of the DSLC series, which served as the source of data for our corpus, and Section 4 describes the dataset construction process and presents statistical information. In Section 5, we discuss the analyses conducted on this corpus and the corresponding results. A brief summary is provided in Section 6, followed by a discussion of limitations and ethical considerations.

## 2 Related Work

To the best of our knowledge, there are few existing datasets of dialogues between humans and multimodal dialogue systems. However, several datasets are available for multimodal dialogues between humans or between humans and a WoZ system.

For example, MELD (Poria et al., 2018) includes video data from the TV series "Friends", with annotations for emotions. Additionally, multimodal datasets specifically focused on emotions, such as IEMOCAP (Busso et al., 2008) and MOSI (Zadeh et al., 2016), have been constructed. The D64 Multimodal Conversational Corpus (Oertel et al., 2013) contains data from natural conversations between humans collected using cameras and motion capture devices. AMI (Kraaij et al., 2005) is a corpus of meetings that includes video recordings of discussions. CEJC (Koiso et al., 2022) captures everyday conversations and contains video data of human-to-human interactions across various daily activities.

One notable multimodal dialogue dataset between humans and a WoZ system is the Hazumi corpus (Komatani and Okada, 2021). This corpus contains casual conversations between humans and a multimodal dialogue system operated by a wizard, along with subjective evaluation scores from users. Analyses examining the relationship between system behavior and user evaluations have been conducted (Wei et al., 2021). However, since the corpus does not include dialogues between humans and autonomous dialogue systems, it is limited in addressing the challenges associated with developing and improving real multimodal dialogue systems.

It should be noted that the term "multimodal dialogue" can also refer to dialogues discussing visual or video contents. Well-known examples include MMConv (Liao et al., 2021), VideoChat (Li et al., 2023), and SIMMC (Kottur et al., 2021). However, these deal with text-based chat systems that interact with images, videos, or virtual reality environments, and do not involve face-to-face interactions typical of human dialogues, which are the focus of this study.

## 3 Dialogue System Live Competitions

The DSLC is a competition for dialogue systems that has been held in Japan since 2018 (Higashinaka et al., 2021). DSLC consists of preliminary and final rounds, with the final round featuring a live event where dialogue systems are demonstrated in front of an audience, and rankings are determined based on audience evaluations. Initially, the competition focused solely on text-based dialogue systems, but starting with DSLC5 in 2022, it expanded to include multimodal dialogues (Higashinaka et al., 2024). In this section, we describe DSLC5 and DSLC6, from which the data for our corpus were sourced. Since the dataset is created from the preliminary round data, the final round is not discussed in this paper.

### 3.1 DSLC5

Two tracks were held in DSLC5: the Open Track and the Situation Track. In the Open Track, systems competed based on their performance in open-domain casual conversation. In the Situation Track, systems were evaluated on their ability to engage in human-like interactions according to predefined scenarios. The specific situation was as follows:

> Shizuka (the system) and Yuki (the user) are friends from the same university seminar group. Shizuka has lost an expensive technical book borrowed from Yuki and is now unable to return it. Shizuka explains the situation and offers an apology to Yuki.

The developers created systems capable of engaging in human-like conversations within this scenario. The scenario was designed based on the Oral Proficiency Interview used in language education, facilitating an effective assessment of language proficiency. In both tracks, the dialogue duration was set to 4 minutes.

Initially, 11 teams entered the Open Track and 15 teams participated in the Situation Track. Due to challenges in system development, several teams withdrew, leaving nine teams in the Open Track and ten teams in the Situation Track for the preliminary round. In both tracks, participants utilized software provided by the organizers to develop their dialogue systems. This software allowed participants to focus solely on implementing the dialogue control module, which sent commands for gestures and facial expressions, along with utterance content, based on the received speech recognition results. The system's gestures and facial expressions were displayed through a CG character named CGErica (Glas et al., 2016), included with the software. In DSLC5, the gestures and facial expressions of users were not processed by the system.

In the preliminary round, a total of ten systems in the Open Track and 11 systems in the Situation Track, including the baseline system for each track, were evaluated. Each system was subjectively assessed by approximately 50 crowdworkers. The systems were operated in the cloud, and dialogues were conducted via Zoom. Since separate crowdworkers were recruited for each system, the evaluators varied across systems.

In the Open Track, the dialogue participants selected two topics from a list of pre-determined keywords and engaged in casual conversation about those topics. The evaluation was based on two aspects: dialogue content and manner of speaking. Dialogue content was assessed using three criteria: Naturalness (whether the dialogue felt natural), Topic Following (whether the system appropriately responded to the chosen topics), and Topic Provision (whether the system could provide new information related to the chosen topics). Each criterion was rated on a 5-point Likert scale. The manner of speaking was evaluated for Naturalness of Interaction (whether the system's manner of speaking, including voice, gestures, and facial expressions, appeared natural), also on a 5-point Likert scale.

In the Situation Track, the systems were comprehensively evaluated based on a single criterion: "How appropriate and human-like was the conversation for the given situation?" (Overall), using a 5-point Likert scale.

### 3.2 DSLC6

In DSLC6, only the Situation Track was conducted. This decision was made because, with the advancement of large language models, sustaining casual conversation was deemed trivial (Iizuka et al., 2023). The software utilized was the same as in DSLC5 but with additional inputs, including estimated emotions (e.g., happy, surprised), head orientation, age, and gender, which were automatically inferred from the user's facial images using open-source software. The situation for the competition was set as follows:

> Yuki (the user), a member of the film club, is considering organizing a welcome party for Professor Kobayashi, who has recently taken over as the club's advisor this month. Yuki consults with Shizuka (the system), a fellow member of the film club, about the plans for the welcome party.

There were initially ten team entries, but due to technical difficulties, only eight teams ultimately participated in the preliminary round. The dialogue duration was set to 5 minutes.

To ensure the accurate capture of users' facial images and other inputs, the preliminary round was conducted in a laboratory setting where evaluators engaged in face-to-face conversations with the systems, rather than using a cloud-based format. Three baseline systems were provided (two using GPT-3.5 with different prompts and one using GPT-4), resulting in a total of 11 systems evaluated. Each system was assessed by approximately 50 evaluators. If a system failed to operate, participants interacted with one of the baseline systems.

The systems were evaluated based on three criteria, each rated on a 5-point scale: Utterance Content (whether the system's responses were contextually appropriate), Gesture/Facial Expression (whether the system's gestures and facial expressions were contextually appropriate), and Voice (whether the system used appropriate timing, tone, and intensity in its speech). Notably, many of the systems in DSLC6 were built on OpenAI's GPT-4 or GPT-3.5 APIs, marking a significant difference from DSLC5, in which rules and locally fine-tuned language models were observed.

## 4 Corpus

As the organizers of the DSLC5 and DSLC6, we processed the data of the preliminary rounds to create a multimodal dialogue corpus between users and multimodal dialogue systems. Specifically, we extracted video segments for each dialogue session, transcribed the audio (transcriptions were performed only for user utterances, as the system's utterances were logged), and linked these with system logs and subjective evaluation scores. The data for each dialogue session consists of an ELAN

| | LC5O | LC5S | LC6S |
|---|---|---|---|
| # System | 10 | 11 | 11 |
| # Dialogue | 537 | 569 | 641 |
| # System Utterance | 13,111 | 17,730 | 20,963 |
| # User Utterance | 11,176 | 12,167 | 15,114 |
| # Total Utterance | 24,287 | 29,897 | 36,077 |
| # Words / System Utt | 10.13 | 11.31 | 11.18 |
| # Words / User Utt | 10.13 | 6.45 | 10.77 |
| Duration (hours) | 42.32 | 43.91 | 57.09 |

Table 1: Statistics of DSLCMM and its subsets. “Utt” stands for “utterance”.

| Subset | Criteria | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| LC5O | Nat | 18.8 | 28.1 | 18.8 | 27.7 | 6.5 |
| | Topic F | 18.2 | 29.6 | 20.5 | 25.3 | 6.3 |
| | Topic P | 19.2 | 28.3 | 19.4 | 27.2 | 6.0 |
| | Nat in Int | 8.9 | 20.7 | 25.0 | 36.5 | 8.9 |
| LC5S | Overall | 4.6 | 12.5 | 18.3 | 48.7 | 16.0 |
| LC6S | Utt Cont | 5.5 | 16.5 | 17.0 | 44.5 | 16.5 |
| | Gest/Face | 5.6 | 21.7 | 23.9 | 39.0 | 9.8 |
| | Voice | 6.9 | 19.7 | 26.8 | 36.7 | 10.0 |

Table 2: Distribution of evaluation scores (%). A score of 1 represents the worst rating, and a score of 5 represents the best. “Nat,” “Topic F,” “Topic P,” and “Nat in Int” refer to Naturalness, Topic Following, Topic Provision, and Naturalness in Interaction, respectively. “Utt Cont” and “Gest/Face” denote utterance content and gesture/facial expression, respectively.

file containing transcriptions, an MP4 video file with recordings (user videos are available only for DSLC6), separate audio files for each speaker, and a JSON file containing system logs and subjective evaluation scores. As a result, we constructed the DSLC Multimodal Corpus (DSLCMM), comprising 1,747 multimodal dialogues.

The overview of the dataset is presented in Table 1. We refer to the subset of the Open Track from DSLC5 as LC5O, and the subset of the Situation Track from DSLC5 as LC5S. The subset from DSLC6 is referred to as LC6S. Each of these subsets contains more than 500 dialogues. The dataset is deemed sufficiently large for analysis and post-training tasks. For detailed statistical information on the systems in each subset, please refer to the appendix.

The distribution of subjective evaluation scores for each subset is presented in Table 2. As shown, the dataset includes a variety of evaluations, reflecting both appropriate and inappropriate dialogue examples. This indicates that the dataset covers a wide range of phenomena observed in multimodal dialogues with systems. Moreover, the relatively small number of instances with the highest evaluation score suggests that there is still room for improvement in the systems.

## 5 Analyses

To illustrate how this corpus can be utilized, we analyzed the relationship between various features of the dialogues and the users’ subjective evaluations (Table 3). Specifically, we extracted features such as the number of utterances, gestures, and facial expressions from each dialogue, and calculated Spearman’s correlations between these features and the subjective evaluation scores. In this context, “gesture” and “face” refer to the number of commands issued for gesture and facial expression outputs, respectively. The logs allowed us to accurately count the number of gesture and facial expression commands. “Latency” denotes the time between the end of the user’s speech and the start of the system’s response.

In LC5O, significant positive correlations were observed between the number of system utterances, the number of gesture and facial expression commands, and the evaluation scores. This suggests that the system made a good impression on users by providing informative utterances and expressing gestures. Regarding Latency, a positive correlation was observed with Topic Following. This indicates that longer response times were associated with better subjective evaluations, potentially because longer response times led to higher-quality responses regarding topics. In LC5S, a significant negative correlation was observed between the number of facial expression commands per utterance and the evaluation score. This suggests that expressing appropriate facial expressions to match specific situations may be more challenging compared to the Open Track.

In LC6S, a negative correlation was found with the number of user utterances, suggesting that systems requiring users to speak extensively were likely more difficult to interact with. Similar to LC5O, the number of gesture commands showed a positive correlation. Additionally, the evaluation score exhibited a positive correlation with the number of facial expression commands. Here, an inverse correlation relative to the LC5S results was observed, suggesting that further detailed analyses of the specific types of expressions and their contextual circumstances are needed.

The analysis presented here is based on overall trends observed across multiple systems, and such insights could not be obtained from a dataset featur-

| Dialogue | LC5O | | | | LC5S | LC6S | | |
|---|---|---|---|---|---|---|---|---|
| Features | Nat | Topic F | Topic P | Nat in Int | Overall | Utt Cont | Gest/Face | Voice |
| # User Utterance | –0.03 | –0.06 | –0.00 | –0.07 | 0.02 | –0.17* | –0.15* | –0.24* |
| # System Utterance | 0.17* | 0.20* | 0.19* | 0.15* | 0.03 | 0.04 | –0.04 | 0.00 |
| # Gesture | 0.13* | 0.14* | 0.08 | 0.11 | –0.01 | 0.23* | 0.12* | 0.17* |
| # Face | 0.18* | 0.18* | 0.15* | 0.20* | –0.10 | 0.13* | 0.01 | 0.02 |
| # Gesture / Utterance | –0.01 | 0.01 | –0.07 | –0.01 | –0.02 | 0.21* | 0.14* | 0.14* |
| # Face / Utterance | 0.04 | 0.04 | 0.03 | 0.08 | –0.12* | 0.11* | 0.07 | 0.04 |
| Latency | 0.11 | 0.17* | 0.02 | 0.03 | 0.04 | –0.05 | 0.00 | 0.01 |

Table 3: Correlation between evaluation scores and features. Asterisks indicate statistical significance ($p < 0.01$).

ing only a single system. DSLCMM enables this type of cross-system analysis, offering valuable and generalizable knowledge. We believe this dataset will be instrumental in advancing understanding in the field.

## 6 Summary

In this study, we utilized data from the DSLC series to construct DSLCMM, a multimodal human-machine dialogue corpus. DSLCMM encompasses dialogues from multiple systems, capturing a wide range of dialogue quality. It includes data on system gestures and facial expression commands, as well as video data featuring facial information of users, making it well-suited for tasks such as multimodal dialogue generation. With user evaluation scores included, the dataset can also support tasks like estimating user evaluation scores (Komatani et al., 2023), detecting dialogue breakdowns in multimodal settings (Higashinaka et al., 2016; Tsubokura et al., 2022; Miah et al., 2024). It can also be useful for building multimodal models for face-to-face conversation (Park et al., 2024; Zhu et al., 2024). We hope this corpus will advance research and development in multimodal dialogue systems.

## Limitations

The dataset constructed in this study is valuable as it contains dialogues between humans and multimodal dialogue systems, but it has certain limitations. The software used by participants is uniform, and there are only two situational contexts, which may limit the variability in dialogues. Additionally, the dialogues are constrained by the capabilities of the systems at the time; similar dialogues might not be generated with faster and more advanced large language models in the future. Furthermore, since the dataset is in Japanese, it is uncertain whether the insights gained here can be applied to other languages.

## Ethical Considerations

The dataset constructed in this study includes users' speech and facial images, necessitating careful consideration of privacy. We have obtained approval from the ethical review committee for departments at the Higashiyama Campus, Nagoya University, concerning data collection, usage, and publication. In releasing this dataset, we will ensure that privacy is rigorously protected, and any data that poses a privacy concern will be promptly withdrawn. There is a potential risk that the data could be used to build dialogue systems that impersonate specific individuals. To address this, we plan to include provisions in the terms of use explicitly prohibiting such applications.

## Acknowledgments

This work was supported by JST Moonshot R&D, Grant number JPMJMS2011. The DSLC events were hosted by the special interest group on Spoken Language Understanding and Dialogue processing (SIG-SLUD) of the Japanese Society for Artificial Intelligence. The events were also supported by JSPS KAKENHI Grant Number 19H05692. We also thank the participating teams of DSLC5 and DSLC6.

## References

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

Dylan F Glas, Takashi Minato, Carlos T Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. 2016. ERICA: The ERATO intelligent conversational android. In *Proc. RO-MAN*, pages 22–29.

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and

Reina Akama. 2021. Dialogue system live competition: identifying problems with dialogue systems through live event. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 185–199.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proc. LREC*, pages 3146–3150.

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

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

Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. 2023. Clarifying the dialogue-level performance of GPT-3.5 and GPT-4 in task-oriented and non-task-oriented dialogue systems. In *Proc. the AAAI Symposium Series*, volume 2, pages 182–186.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the corpus of everyday Japanese conversation. In *Proc. LREC*, pages 5587–5594.

Kazunori Komatani and Shogo Okada. 2021. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *Proc. ACII*, pages 1–8.

Kazunori Komatani, Ryu Takeda, and Shogo Okada. 2023. Analyzing differences in subjective annotations by participants and third-party annotators in multimodal dialogue corpus. In *Proc. SIGDIAL*, pages 104–113.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. 2024. PlatoLM: Teaching LLMs in Multi-Round Dialogue via a User Simulator. In *Proc. ACL (Volume 1: Long Papers)*, pages 7841–7863.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The AMI meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: an environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 675–684.

Md Messal Monem Miah, Ulie Schnaithmann, Arushi Raghuvanshi, and Youngseo Son. 2024. Multimodal contextual dialogue breakdown detection for conversational ai models. *arXiv preprint arXiv:2404.08156*.

Catharine Oertel, Fred Cummins, Jens Edlund, Petra Wagner, and Nick Campbell. 2013. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7(1):19–28.

Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. 2024. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *Proc. ACL (Volume 1: Long Papers)*, pages 16334–16348.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Kazuya Tsubokura, Yurie Iribe, and Norihide Kitaoka. 2022. Dialog breakdown detection using multimodal features for non-task-oriented dialog systems. In *Proc. GCCE*, pages 352–356.

Wenqing Wei, Sixia Li, Shogo Okada, and Kazunori Komatani. 2021. Multimodal user satisfaction recognition for non-task oriented dialogue systems. In *Proc. ICMI*, pages 586–594.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proc. LREC*, pages 1556–1559.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. 2024. INFP: Audio-driven interactive head generation in dyadic conversations. *arXiv preprint arXiv:2412.04037*.

# A Appendix

The following tables present the statistical data for each system within the subsets of DSLCMM. The terms “# Utt”, “# Gesture”, and “# Face” refer to the number of utterances per dialogue, as well as system commands per dialogue for gestures and facial expressions, respectively.

| System | # Dialogue | # Utterance | | # Vocabulary | | System | | | User |
|---|---|---|---|---|---|---|---|---|---|
| | | System | User | System | User | # Utt | # Gesture | # Face | # Utt |
| LIO | 50 | 2301 | 1245 | 1814 | 1174 | 46.02 | 15.32 | 4.00 | 24.90 |
| TOH | 51 | 1036 | 979 | 2085 | 1209 | 20.31 | 357.73 | 1317.55 | 19.20 |
| BAO | 72 | 1447 | 1650 | 1910 | 1654 | 20.10 | 0.17 | 58.58 | 22.92 |
| TOA | 59 | 2472 | 1078 | 1254 | 1184 | 41.90 | 82.12 | 39.05 | 18.27 |
| AO1 | 51 | 909 | 998 | 1114 | 1167 | 17.82 | 110.75 | 91.18 | 19.57 |
| MIN | 55 | 953 | 1068 | 1204 | 1291 | 17.33 | 103.53 | 194.22 | 19.42 |
| CHU | 49 | 1049 | 1057 | 1011 | 1086 | 21.41 | 52.65 | 198.94 | 21.57 |
| IRI | 44 | 868 | 888 | 806 | 1057 | 19.73 | 4.00 | 6.14 | 20.18 |
| AO2 | 53 | 946 | 953 | 1241 | 1077 | 17.85 | 5.17 | 37.75 | 17.98 |
| AO3 | 53 | 1130 | 1260 | 1336 | 1230 | 21.32 | 21.79 | 0.00 | 23.77 |

Table 4: Statistics of dialogues for systems in LC5O.

| System | # Dialogue | # Utterance | | # Vocabulary | | System | | | User |
|---|---|---|---|---|---|---|---|---|---|
| | | System | User | System | User | # Utt | # Gesture | # Face | # Utt |
| FCL | 54 | 1704 | 1094 | 173 | 534 | 31.56 | 17.83 | 34.89 | 20.26 |
| LIS | 54 | 1107 | 1227 | 527 | 622 | 20.50 | 36.80 | 17.15 | 22.72 |
| YUR | 50 | 2683 | 752 | 294 | 408 | 53.66 | 97.62 | 50.20 | 15.04 |
| NAK | 58 | 2433 | 1340 | 182 | 642 | 41.95 | 34.47 | 47.10 | 23.10 |
| AS1 | 52 | 1245 | 1074 | 180 | 590 | 23.94 | 8.58 | 28.65 | 20.65 |
| BAS | 48 | 1764 | 1099 | 208 | 621 | 36.75 | 6.79 | 5.00 | 22.90 |
| CIT | 52 | 1361 | 1230 | 187 | 622 | 26.17 | 127.50 | 95.29 | 23.65 |
| SAI | 51 | 1619 | 1376 | 117 | 660 | 31.75 | 6.04 | 16.00 | 26.98 |
| HON | 53 | 953 | 1102 | 188 | 561 | 17.98 | 35.17 | 119.25 | 20.79 |
| AS2 | 43 | 1477 | 1039 | 245 | 560 | 34.35 | 1.93 | 25.19 | 24.16 |
| TSU | 54 | 1384 | 834 | 263 | 476 | 25.63 | 87.26 | 649.39 | 15.44 |

Table 5: Statistics of dialogues for systems in LC5S.

| System | # Dialogue | # Utterance | | # Vocabulary | | System | | | User |
|---|---|---|---|---|---|---|---|---|---|
| | | System | User | System | User | # Utt | # Gesture | # Face | # Utt |
| TOH | 50 | 1330 | 1026 | 530 | 881 | 26.60 | 184.40 | 61.30 | 20.52 |
| BI3 | 105 | 2036 | 2572 | 1064 | 1427 | 19.39 | 470.47 | 29.12 | 24.50 |
| RIS | 51 | 3351 | 1284 | 654 | 911 | 65.71 | 563.37 | 59.39 | 25.18 |
| UEC | 58 | 3192 | 1270 | 1012 | 1031 | 55.03 | 492.28 | 82.47 | 21.90 |
| BI4 | 56 | 1013 | 1276 | 848 | 1070 | 18.09 | 483.46 | 26.79 | 22.79 |
| BK3 | 54 | 1109 | 1381 | 998 | 1027 | 20.54 | 516.81 | 31.43 | 25.57 |
| HNL | 53 | 1819 | 1293 | 1026 | 948 | 34.32 | 373.23 | 10.13 | 24.40 |
| YAM | 50 | 2271 | 1534 | 718 | 1015 | 45.42 | 152.56 | 22.48 | 30.68 |
| CIT | 56 | 1414 | 1139 | 757 | 943 | 25.25 | 174.27 | 132.86 | 20.34 |
| AN1 | 50 | 2546 | 1116 | 1220 | 1024 | 50.92 | 125.98 | 46.88 | 22.32 |
| AN2 | 58 | 882 | 1223 | 982 | 967 | 15.21 | 3.12 | 12.19 | 21.09 |

Table 6: Statistics of dialogues for systems in LC6S.