Interaction Matters: An Evaluation Framework for Interactive Dialogue Assessment on English Second Language Conversations

Rena Gao and **Carsten Roever** and **Jey Han Lau**
University of Melbourne, Australia
wegao@student.unimelb.edu.au, {carsten, laujh}@unimelb.edu.au

Abstract

We present an evaluation framework for interactive dialogue assessment in the context of English as a Second Language (ESL) speakers. Our framework collects dialogue-level interactivity labels (e.g., topic management; 4 labels in total) and micro-level span features (e.g., backchannels; 17 features in total). Given our annotated data, we study how the micro-level features influence the (higher level) interactivity quality of ESL dialogues by constructing various machine learning-based models. Our results demonstrate that certain micro-level features strongly correlate with interactivity quality, like reference word (e.g., she, her, he), revealing new insights about the interaction between higher-level dialogue quality and lower-level fundamental linguistic signals. Our framework also provides a means to assess ESL communication, which is useful for language assessment[1].

1 Introduction

Estimates suggest more than 750 million individuals use English as a non-native language (Dyvik, 2023). Despite its widespread use, a notable gap exists in the availability of datasets that capture the communicative features of English Second Language (ESL) speakers within dialogic contexts. Most existing dialogue datasets are primarily created with native speakers' conversations, failing to consider the distinct linguistic subtleties and obstacles encountered by ESL speakers (Settles et al., 2021) such as different usages on grammar, syntax and sentence structure influenced by their native languages. On the other hand, for dialogue quality evaluation, most existing performance metrics focus on fluency, coherence or consistency (Tao et al., 2018), which fail to capture or evaluate the sophisticated features of dialogue such as the speakers' ability to interact, manage topics through multi-turn dialogues, or use the appropriate tone given a particular domain/context. These gaps, in particular, are becoming more crucial due to the increasing demand to evaluate ESL speakers' communication and interaction skills, which is important not only for better cross-cultural exchanges but also for improving educational assessments. While resources such as the International Corpus of Learner English (Rica-Peromingo, 2009) offer data from controlled spoken settings on monologic speech, they fall short in addressing multi-party interactive dialogues.

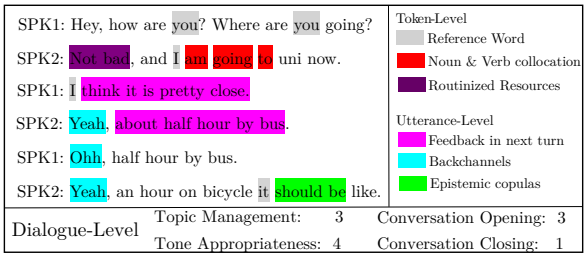


Figure 1: An example of an annotated dialogue with dialogue-level interactivity labels and micro-level features

In this paper, we introduce an ESL dialogue dataset and propose an evaluation framework designed to capture dialogue interactivity. Specifically, our framework has two different levels of annotation: (1) 4 dialogue-level interactivity labels that capture topic management, tone appropriateness and conversation opening and closing; and (2) 17 micro-level linguistic features that capture token-level features (e.g., reference word and routinized resources) and utterance-level features (e.g., epistemic copulas and backchannels). Figure 1 illustrates an example of an annotated dialogue. Appendix A.3 gives the full list of interactivity labels and micro-level features, along with their descriptions. Note that the micro-level features are annotated as spans, while the dialogue interactivity

[1]The dataset and code are available at: https://github.com/RenaGao/2024InteractiveMetrics

Proceedings of the 31st International Conference on Computational Linguistics, pages 10977–11012
January 19–24, 2025.

labels are document labels.

After annotating the ESL dialogues with our framework, we investigate the relationship between interactivity labels and micro-level features. To this end, we build machine learning models that use micro-level features as input to predict the interactivity labels of a dialogue. We demonstrate how micro-level features impact various interactive aspects of ESL dialogues: specifically we saw which micro-level features contribute to the prediction of a particular interactivity quality. Additionally, we also compare against a baseline BERT (Devlin et al., 2018) that uses the *raw dialogue* as input to predict the interactivity labels, and found that it performs worse than our (simpler) machine learning models that use micro-level features as input, suggesting that these micro-level features have a stronger predictive power for interactivity. To summarize, our contributions are given as follows:

- We propose a novel evaluation framework for ESL dialogues that assesses four dialogue-level interactivity labels, including topic management, tone appropriateness and conversation opening and closing. It also captures seventeen fundamental micro-level features, such as backchannels (at the utterance-level) and reference words (at the token-level).
- We release SLEDE (**S**econd **L**anguage **E**nglish **D**ialogue **E**valuation), an annotated ESL dialogue dataset based on our evaluation framework.
- We study the interplay between the interactivity labels and micro-level features via predictive learning. Our experimental results explain how certain micro-level features impact various interactive aspects of ESL dialogues. Our predictive models have the potential to be applied to real-world English tests to assess ESL communication.

2 Related work

2.1 ESL Conversational Dialogue

The interactive feature of human dialogue influences how turns and overlaps occur when analyzing conversations in communication, which is important for tagging and processing dialogue data (Allwood, 2008). Due to the complex nature of data collection and practical issues, open-source conversational dialogues are still limited in related research fields, and most conversational datasets are designed for speech recognition purposes (Lovenia et al., 2022). The interactive feature of conversations vary between English native speakers and ESL speakers. For native speakers, the fluidity and nuance of the language come naturally, allowing for a dynamic range of expressions and a deeper level of engagement in conversation. However, ESL speakers often navigate different social and cultural norms through the usage of a second language, which adds complexity and richness to the conversation dataset and reflects the multifaceted nature of human communication. Moreover, the learners' native languages frequently shape their learning and usage of a second language, resulting in distinct constructions, mistakes, and use patterns (Betts, 2003; Warren, 2017). As a consequence, it is interesting to ask the following questions when creating a second language conversation dataset: (1) how can we annotate lower level grammar related and communicative features?; and (2) how can we capture the higher level dialogue interactivity qualities?

2.2 Dialogue Interactivity Quality

Our evaluation framework assesses on four interactivity quality in dialogue: topic management, tone appropriateness, and conversation opening and closing. Here we discuss various studies focusing on these aspects, providing motivation on why we choose them in our framework.

Topic Management How speakers collaboratively manage topics in a dialogue is an important indicator of interactional ability. Speakers exhibit increasing mutuality and engagement in their interactions (Galaczi, 2014). They demonstrate mutuality by taking up and extending interlocutor-initiated topics through reformulating interlocutor contributions (Lam, 2018), and they provide frequent listener responses and assessments of interlocutor statements ("that's so cool", "definitely", "oh no"), thus creating a stronger sense of engagement (Galaczi, 2014). Ghazarian et al. (2022) argued that evaluating topic coherence in human conversation is still a challenging task and called for a more empirical way of conducting this evaluation.

Tone Appropriateness Whittaker et al. (2021) suggested the social role of a chatbot needs to be emphasised when measuring chatbot performance. As such, another important aspect of interactional ability is language choice following the social role. Pill (2016) demonstrated the need for healthcare

professionals to speak at a high level of linguistic proficiency and speak in ways particular to their profession. Dai (2022) and Dai and Davey (2023) extended this work to other social roles and showed that language users are capable of configuring their linguistic abilities to display attributes commonly associated with a particular social role in their interactions. Roever and Dai (2021) and Roever and Ikeda (2023) similarly found that humans learn to talk in ways conventionally expected for a social role.

Conversation Opening and Closing Opening and closing of conversations is a long-standing fundamental research concern in dialogues (Schegloff, 1968; Schegloff and Sacks, 1973), which can also be used to differentiate levels of interactional ability. Proficient ESL speakers are found more likely to open the conversation with preliminary and affiliative talk than less proficient speakers (Abe and Roever, 2019). Similarly, proficient ESL speakers are shown to display more elaborate closings (Abe and Roever, 2020). Stolcke et al. (2000), however, argued that there is still a lack of practical measures to assess the performance of starting and closing a conversation.

2.3 Dialogue Fundamental Features

There is a growing interest in developing more sophisticated evaluation frameworks that can adapt to the diverse grammatical structure of spoken interaction in dialogues (Sinha et al., 2020), which is essential for understanding ESL communication. Dinan et al. (2020) argued that more advanced semantic analysis tools are needed to better understand vocabulary choices' impact on dialogue quality from a micro-level, including code-switching, response tokens, and tense choice for verbs. Currently, only limited works have discussed the empercial methods on how to link these vocabulary choices to demonstrate the quality of communication in conversations.

For a bigger unit, utterance level features such as feedback in next turn and backchannels are all critical features in considering the quality of interactions (Wu and Roever, 2021). In addition, notion in grammatical resources, such as modal verbs (Shaxobiddin, 2024), epistemic copulas (Hayashi, 2020), and collaborative finishes (Yap and Sahoo, 2024), highlight the ability in deploying basic fundamental resources in actual interaction when constructing a dialogue. Thus, the evaluation metrics need to be sensitive to the linguistic features of multiple languages and the contexts in which these choices occur. The 17 micro-level features in our framework are inspired by these studies.

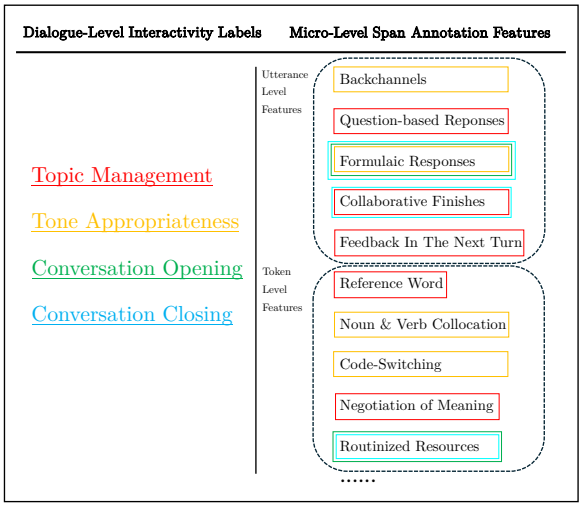


Figure 2: Our proposed evaluation framework has dialogue-level interactivity labels and micro-level features targeting interaction and engagement.

3 Evaluation Framework

Our motivation is to design a more comprehensive and transparent dialogue evaluation framework that captures dialogue interactivity and fundamental linguistics properties. To this end, we introduce an evaluation framework that has two levels of annotations: (1) dialogue-level interactivity labels (4 labels); and (2) micro-level linguistic features (17 features).

For the interactivity labels, we annotate: (1) topic management, which measures how extensively the topic is expanded upon and whether the content is new or previously discussed; (2) tone appropriateness, which indicates the degree of formality; (3) conversation opening, which rates the quality of greetings and (4) conversation closing, which rates the quality of summaries. Each of these labels is annotated with five categorical scores from 1 to 5 to assess the degree of interactivity; Table 1 provides a detailed description for each score.

For micro-level features, we target grammatical, interactional and semantic aspects, and further decompose them into 7 token-level features that represent word formations that are indicative of an ESL speaker's ability to navigate linguistic resources for clarity, emphasis, and cultural relevance, including "reference word", "noun & verb collocation in proper form", "code-switching for

Interactivity Labels	Scores	Description of Scores
Topic Management	[5]	topic extension with clear new context
	[4]	topic extension under the previous direction
	[3]	topic extension with the same content
	[2]	repeat and no topic extension
	[1]	no topic extension and stop the topic at this point
Tone Appropriateness	[5]	very informal
	[4]	quite informal, but some expressions are still formal
	[3]	relatively not formal, and most expressions are quite informal
	[2]	quite formal, and some expressions are not that formal
	[1]	very formal
Conversation Opening	[5]	nice greeting and showing a good understanding of the opening of conversation in social interactions.
	[4]	sounded greeting and showed a basic understanding of the social role.
	[3]	general greeting but not understanding the social role well.
	[2]	basic greeting.
	[1]	no opening, start the discussion immediately.
Conversation Closing	[5]	detailed summarization and smooth transition to the closing of the conversation.
	[4]	transit to the closing naturally, but without summarising the discussion.
	[3]	transit to the discussion.
	[2]	demonstrate a translation to the end of the conversation.
	[1]	no closing, directly stop the conversation.

Table 1: Description of scores for dialogue-level interactivity labels. Higher score indicates better interactivity ability, for example, *Tone Appropriateness* scores higher with more informality shows that the speakers are able to employ more active linguistics resources in dialogue communication to perform more causal and natural interactions compared with the formal tone, which has limited linguistics resources and not naturally occurred in real-life conversations.

communicative purposes", "negotiation of meaning", "tense choice to indicate interactive aims", "routinized resources" and "subordinate clauses" ; and 10 utterance-level features for contextual interactions including "backchannels"; "responses framed as questions"; "formulaic expressions"; "collaborative finishes"; "adjectives and adverbs denoting possibility"; "constructions with impersonal subjects" followed by "non-factive verbs and noun phrases" and "feedback in the next turn", "impersonal subject + non-factive verb + NP", "adjectives/ adverbs expressing possibility". These features capture the dynamic interplay between speakers, emphasizing the importance of backchannels, question-framed responses, and other mechanisms that facilitate a collaborative and adaptive exchange. Figure 2 summarises our evaluation framework, and Appendix A.3 provides the full details of these labels/features.

4 SLEDE Development

We now describe how we develop SLEDE (Second Language English Dialogue Evaluation): i.e. how we collect ESL dialogue data (Section 4.1) and annotate the data based on our evaluation framework (Section 4.2).

4.1 ESL Dialogue Collection

We first look at finding the right set of conversational topics for the participants. We came up with a preliminary set of topics, and survey a group of 60 individuals, comprising both native English speakers and ESL speakers, to get their feedback on the quality of the topics. After collecting the feedback, we used their insights to further refine the topic set; the final set of topics are presented as part of the questionnaire that we ask participants to fill in before we collect their dialogues ("block 4" in Appendix A.4).

Next, we recruit 120 Chinese ESL speakers (volunteers) to engage in a 1-to-1 in-person talk on a chosen topic. The criteria for selecting volunteers for collecting the datasets are given as follows: (1) An IELTS score exceeding 6.5 to comprehend the dialogue fully; (2) A minimum educational attainment of a bachelor's degree in data science, computer science, or linguistics from a recognized university; (3) Consent to agree on recording (refer to Appendix A.5 for details). These prerequisites were established to guarantee that the workers possess proficient English comprehension and are adequately equipped to have a high-quality conversation for the pair discussion. All speakers will then go through a training phase to ensure they un-

Datasets	Full dialogues
# dialogues	120
# turns (max)	2,065
# turns (avg)	1,760
# words marked (token-level features)	10,852
# words marked (utterance-level features)	3,516
# micro-level features (total counts)	14,386

Table 2: Annotated Data Statistics

Measure	Token-level Features	Utterance-level Features	Dialogue-level Labels
α	0.63	0.64	0.65
r	0.64	0.67	0.68

Table 3: Inter-annotator agreement for micro-level features (token-level and utterance-level) and dialogue-level labels.

derstand the task. We follow Mehri et al. (2022) where we provide instructions (Appendix A.5) to highlight important dialogue aspects to take into account, such as coherence, language complexity, and naturalness.

After training, we break the 120 volunteers into 60 pairs. The pair was matched with similar second language proficiency to ensure the dialogue maintains a stable quality within the two speaker's interaction, for example, a IELTS 6 ESL speaker was paired with another IELTS 6.5 speaker. Each pair undergoes two rounds of conversation: the first half-hour is dedicated to discussing a specific topic (chosen by them in the questionnaire), and the second half-hour involves discussing a specific issue and proposing solutions (Appendix A.5). We therefore collected a total of 120 dialogues, each lasting about half an hour, with thousands of turns in each dialogue.

4.2 Dialogue-level Interactivity Label and Micro-level Feature Annotation

Given the 120 dialogues, we now collect annotations based on our proposed evaluation framework (Section 3). To this end, we recruit eleven volunteer postgraduate students proficient in English (six in computer science and four in applied linguistics). These eleven annotators and the first author were randomly split into six pairs to annotate the dialogues.

The annotators were presented with an annotation guide (Appendix A.3) to explain the dialogue-level interactivity labels and micro-level features. It includes definitions and examples of each label/feature, as well as guidelines for using the annotation interface. For the dialogue-level interactivity labels (topic management, tone appropriateness, conversation opening and closing), the annotators are asked to give a score for each of the four labels, and the task is framed as a document labelling task. We adopt a majority voting approach to annotate labels, e.g., if annotators give different labels to the same dialogue, we select the most frequent label as our final label. For micro-level features, they are framed as a span annotation task where the annotators are asked to mark word spans that exhibit a particular micro-level feature. Note that a word can be marked with multiple features.

To ensure the quality of the annotation process, annotators went through a training process where they were first asked to label six pilot dialogues, and the first author cross-checks all annotations. Any mistakes are then discussed. After the training, each pair of annotators (including the first author) are given 30 dialogues to annotate (noting that there is some overlapping dialogues between pairs). In total, 120 dialogues are annotated; some statistics of the annotated dataset are presented in Table 2.

To understand annotation quality, we compute inter-annotator agreement for the interactivity labels and micro-level features. For the interactivity labels, we compute agreement between the annotators in a pair and take the average across the pairs. For the micro-level features, we again measure agreement between the annotators in a pair at the token-level for each micro-level feature — i.e., we first break the dialogue into individual word tokens and compute statistics based on the presence or absence of the feature as marked by the annotators for each word token[2] — before aggregating over the features and pairs. We calculate Pearson correlation coefficient r (Cohen et al., 2009) and Krippendorff's α (Krippendorff, 2018) to measure inter-annotator agreement, and the results are summarized in Table 3. The agreement is above 0.6 for micro-level features (token-level and utterance-level) and dialogue-level labels, indicating that there is a good consensus among annotators and the evaluation framework is robust/reliable.

[2]In other words, the unit of analysis here is a word token, and the output is a binary value for each annotator indicating whether it has been marked for the feature.

5 Experiments

We conduct a series of experiments to analyse the influence of micro-level features on dialogue-level interactivity labels. To this end, we first build machine learning models to evaluate the prediction performance of interactivity labels given micro-level features as input in Section 5.1, and then analyse the importance of micro-level features in Section 5.2 and lastly look at the difference between utterance-level vs. token-level features in Section 5.3.

Given that our ESL dialogues are very long (maximum of 2065 turns as shown in Table 2) and we only have a small number of them (120 dialogues), we break each dialogue into smaller "mini-dialogues" that have a maximum of 12 turns in our experiments. This process produces 625 mini dialogues in total. For the micro-level labels, we can carry across the annotations we have collected for the original dialogues. For the interactivity labels, however, we copy the original labels from the larger dialogue they belong to. To measure the validity of this approach, we randomly sample 60 mini-dialogues and re-annotate them (with 6 annotators) for the interactivity labels. Then, we measure the correlation between the two judgements (i.e., judgements copied from the original dialogues vs. judgements collected using mini-dialogues). We found the Pearson correlation to be 0.72, suggesting that our approach of copying the interactivity labels from the larger dialogue is a sensible way of creating labels for the mini-dialogues. Henceforth, all experiments that we describe use the mini-dialogues.

5.1 Predicting Dialogue interactivity labels

We experiment with three machine learning algorithms, logistic regression (LR), random forest (RF), and Naïve Bayes (NB), for predicting each dialogue interactivity label using the micro-level features as input. We frame this as a classification problem, where the model needs to output one of the five classes. For each micro-level feature, the feature weight (x) of a mini-dialogue is computed as a weighted average of the fraction of marked tokens over the annotators:

$$x = \sum_{i=1}^{N} \frac{c_i}{\sum_{j=1}^{N} c_j} \times \frac{c_i}{c_{\text{total}}} \quad (1)$$

where N is the number of annotators who worked on the mini-dialogue, c_i the number of marked word tokens by annotator i, and c_{total} the total number of word tokens in the mini-dialogue. Intuitively, we give more weights to annotators who highlight more words than those who highlight less, and the rationale for doing this is that *under-marking* is a type of mistake more prevalent than *over-marking*, based on a preliminary analysis of the data (and so annotators who don't mark many words should be down-weighted, as their annotations are likely to be of lower quality).

Classification Models				
Labels	Topic	Tone	Opening	Closing
Logistic Regression				
ACC	0.815	0.849	0.975	0.950
PRE	0.690	0.746	0.950	0.941
REC	0.815	0.849	0.975	0.950
F1	0.747	0.794	0.962	0.945
Random Forest				
ACC	0.832	0.832	0.966	0.966
PRE	0.714	0.744	0.950	0.934
REC	0.832	0.832	0.966	0.966
F1	0.766	0.786	0.958	0.950
Naïve Bayes				
ACC	0.807	0.840	0.966	0.958
PRE	0.688	0.733	0.950	0.934
REC	0.807	0.840	0.966	0.958
F1	0.743	0.783	0.958	0.946
BERT				
ACC	0.528	0.530	0.719	0.746
PRE	0.519	0.609	0.647	0.682
REC	0.617	0.584	0.708	0.713
F1	0.572	0.620	0.733	0.752

Table 4: The classification prediction results with different performance metrics accuracy (ACC), precision (PRE), recall (REC) and f1 score (F1) on the SLEDE dataset.

We also include a baseline, where we fine-tune BERT using the *raw dialogue* as input to predict the interactivity labels.[3] This baseline tells us whether the micro-level features are actually useful, or we can use the raw dialogues directly for predicting the interactivity labels. We summarize our results in Table 4 over four metrics: accuracy (ACC), precision (PRE), recall (REC), and F1 Score (F1).

From the results, we observe that the three simple models (LR, RF, and NB) perform exceptionally well on conversation opening and closing, often achieving or nearing 0.95 and above for all

[3]We use 'bert-base-uncased'.

LR	RF	NB
Code Switching	**Code Switching**	**Feedback in Next Turn***
Reference Word*	**Feedback in Next Turn***	**Formulaic Responses**
Feedback in Next Turn*	Question-based responses	**Reference Word***
Formulaic Responses	Non-factive Verb	Negotiation of Meaning
Tense Choice	**Reference Word***	**Tense Choice**

Table 5: High impact common micro-level features over the three classifiers for predicting dialogue-level labels. Bold/asterisk indicates overlapping features in two/three classifiers.

Topic	Tone	Opening	Closing
Logistic Regression			
Negotiation of Meaning*	**Routinized Resources***	Epistemic Modals	**Backchannels***
Subordinate Clauses*	Adj./Adv. Expressing	**Formulaic Responses**	**Adj./Adv. Expressing**
Noun&Verb Collocation	**Feedback in Next Turn***	**Question-Based Responses***	Formulaic Responses
Question-Based Responses	**Formulaic Responses***	**Subordinate Clauses***	**Collaborative Finishes***
Negotiation of Meaning	**Reference Word**	**Adj./Adv. Expressing***	Epistemic Copulas
Naïve Bayes			
non-factive verb phrase structure	**Routinized Resources***	**Adj./Adv. Expressing***	**Adj./Adv. Expressing**
Question-Based Responses	**Feedback in Next Turn***	**Routinized Resources**	Epistemic Modals
Adj./Adv. Expressing	**Epistemic Copulas**	**Subordinate Clauses***	**Backchannels***
Negotiation of Meaning*	Question-Based Responses	Epistemic Copulas	**Collaborative Finishes***
Subordinate clauses*	**Subordinate Clauses***	**Question-Based Responses***	Question-Based Responses
Random Forest			
Negotiation of Meaning*	**Epistemic Copulas**	Feedback in Next Turn	Feedback in Next Turn
Formulaic Responses	Backchannels	**Subordinate Clauses***	Subordinate clauses
Subordinate Clauses*	**Feedback in Next Turn***	**Adj./Adv. Expressing***	**Collaborative Finishes***
Epistemic Copulas	**Negotiation of Meaning**	**Question-Based Responses***	**Formulaic Responses**
Question-Based Responses	**Routinized Resources***	**Formulaic Responses**	**Backchannels***

Table 6: High impact interactivity-specific micro-level features. For each interactivity label, bold/asterisk indicates overlapping features in two/three classifiers.

metrics, indicating that these labels are easier to predict as they only appear at the beginning and the end of the conversation. Topic management and tone prediction, on the other hand, has a lower performance, and it is unsurprising given that it is arguably a more difficult task. That said, we're still seeing over 75% F1 performance in most cases, suggesting that the micro-level features predictive of these interactivity labels.

Interestingly, BERT consistently underperforms by a large margin compared to the simple models. This implies the micro-level features are more predictive of the interactivity labels, and the raw dialogue alone does not provide the same level of information and pretraining isn't good enough close the gap.

Looking at the differences between classifiers, we see largely similar/consistent results, suggesting that the predictive performance is agnostic to the exact implementation of the classifier. We want to note that due to the lack of other ESL conversation datasets, these classifiers are trained from scratch (without having any form of pretraining). Compared to previous studies that found poor performance in classifying topics (Stolcke et al., 2000) and tone choices (Ghazarian et al., 2022) our results are encouraging.

Taking all these observations together, given the relatively strong classification performance, the main insight we can draw here is that the micro-level features are able to explain the four dialogue interactivity qualities, shedding light into the possibility of using this interactive framework in the evaluation of dialogue beyond the ESL context. That is, one future direction for developing dialogue evaluation metrics is to consider incorporating some of these micro-level token and utterance features.

5.2 Feature Importance Analysis

In this section, we further examined the significance of token and utterance-level features for predicting dialogue interactivity, aiming to identify the most important linguistic features influencing different interactive perspectives. This approach may provide insights into the foundational elements that drive the dialogue engagement.

Given that a trained LR, NB and RF classifier all provide weights to indicate the importance of each feature, for each classifier we first compute *common* micro-level features f_c across the four interactivity labels:

$$f_c = \text{top5}\big(\text{top10}(f_{\text{topic}}) \cap \text{top10}(f_{\text{tone}}) \cap \text{top10}(f_{\text{opening}}) \cap \text{top10}(f_{\text{closing}})\big)$$

where topk is a function that returns the best k items given by their weights, f_{topic} denote the set of micro-level features with their weights for predicting the topic management interactivity label.

We display these common features in Table 5 for LR, RF and NB. Interestingly, for common Top-5 features across three models, we observed that "Code-Switching", "Reference Word", and "Tense Choice" are shared across all classifiers (asterisk), and "Formulaic Responses", "Code Switching", "Feedback in Next Turn" is common across two out of three classifiers. We see very consistent highly impact common micro-level features over these different classifiers, suggesting that these features are reliable for predicting the interactivity labels. From the perspective of linguistic constructs for interactive purpose, "Reference Word" indicates the proper person to refer to in dialogue construction (Roever and Ikeda, 2023); for "Code-Switching", and "Tense Choice" demonstrate the ability of smoothing the communication for second language speakers. "Feedback in Next Turn" presents the awareness of giving immediate responses in-time, which is essential in ensuring the dialogue quality in second language interaction.

We next look at micro-level features that are specific to each of the interactivity label. To that end, for each classifier we compute interactivity-specific features, e.g., for topic management, as follows:

$$\text{top10}(f_{\text{topic}}) - f_c \quad (2)$$

Results for presented in Table 6. Again, we see consistent results between classifiers for each interactivity label. Many of these interactivity-specific micro-level features are intuitive. For example, for topic management, we have "Negotiation of Meaning" and "Subordinate Clauses" because these micro-level features tell us about the content and discourse of the discussion and indicate the transitions for topics. For tone appropriateness, "Routinized Resources", "Formulaic Responses", and "Feedback in Next Turn" are essential to demonstrate the social role in language resources choice. For conversation opening, "Question-Based Responses","Subordinate Clauses", and "Adj./Adv. Expressing" all show how a dialogue will be started from both speakers. And lastly for conversation closing, "Collaborative Finishes" and "Formulaic Responses" are directly link to the development of how to end a dialogue.

To conclude, the largely consistent results between classifiers suggest that our findings are robust and not sensitive to the implementation of the classifier. That said, the fact there is some (minor) difference does suggest that there is perhaps complementarity between these classifiers and points to a potential future direction of ensembling these classifiers to improve the prediction of interactivity labels.

5.3 Ablation Study

We now examine the individual effects of utterance-level and token-level features in the learning model predictions for the four interactivity qualities. As before, we train three classifiers (LR, RF, and NB) but this time they use only either the token-level ("Token") or utterance-level ("Utt.") features; results are presented in Table 7. Note that we also include the original results using both sets of features ("Both") for comparison. The results indicate that predictions at the token level are better than those at the utterance level, though the difference isn't large. Perhaps most importantly, we see that using both features together produce the best performance in most of the cases (exceptions: RF for Topic and Tone), showing that both types of micro-level features are important for predicting the dialogue-level interactivity labels.

6 Conclusions

In this paper, we propose a novel evaluation framework to assess the dialogue quality of ESL conversations by considering high-level interactivity labels and micro-level linguistic features. We develop SLEDE, an annotated ESL dialogue corpus

Models	Token	Utt.	Both	Token	Utt.	Both
	Topic			Tone		
LR	0.571	0.658	0.747	0.690	0.609	0.794
RF	0.888	0.799	0.766	0.911	0.898	0.786
NB	0.589	0.576	0.743	0.680	0.673	0.783
	Opening			Closing		
LR	0.915	0.840	0.962	0.934	0.711	0.945
RF	0.974	0.978	0.958	0.976	0.981	0.950
NB	0.915	0.914	0.958	0.934	0.928	0.946

Table 7: The F1 results with different machine learning models across different feature levels.

based on the evaluation framework. We found that the micro-level features are highly predictive of the interactivity labels, and revealed impactful micro-level features that are: (1) common across different interactivity labels; and (2) specific to a particular interactivity label. Our results provide new insights into educational assessment for ESL communication.

7 Limitations

The developed dataset is admittedly small (120 dialogues). That said, the quality of the annotation is high (strong annotator agreement) and each dialogue is very long (almost 1,800 turns on average per dialogue). Ultimately, as our goal is to analyse the relationship between micro-level vs. interactivity features, our predictive models do not offer an end-to-end approach for evaluating dialogue quality, as it requires micro-level features as input. The future work will extend the scope by automating the processing for micro-level features.

Ethics Statement

This study is conducted under the guidance of the ACL Code of Ethics. We manually filtered out potentially offensive content and removed all information related to the identification of annotators. The annotation protocol is approved under the The University of Melbourne's Human Ethics Application (reference number The University of Melbourne Human Ethics LNR as 2022-24988-32929-3.).

References

Makoto Abe and Carsten Roever. 2019. Interactional competence in l2 text-chat interactions: First-idea proffering in task openings. *Journal of Pragmatics*, 144:1–14.

Makoto Abe and Carsten Roever. 2020. Task closings in l2 text-chat interactions: A study of l2 interactional competence. *Calico Journal*, 37(1):23–45.

Jens Allwood. 2008. Dimensions of embodied communication-towards a typology of embodied communication. *Embodied communication in humans and machines*, pages 257–284.

Robert Betts. 2003. Easyenglish: Challenges in cross-cultural communication. In *EAMT Workshop: Improving MT through other language technology tools: resources and tools for building MT*.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

David Wei Dai. 2022. *Design and validation of an L2-Chinese interactional competence test*. Ph.D. thesis, University of Melbourne (Australia).

David Wei Dai and Michael Davey. 2023. On the promise of using membership categorization analysis to investigate interactional competence. *Applied Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.

Einar H. Dyvik. 2023. The most spoken languages worldwide 2023.

Evelina D Galaczi. 2014. Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied linguistics*, 35(5):553–574.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. Deam: Dialogue coherence evaluation using amr-based semantic manipulations. *arXiv preprint arXiv:2203.09711*.

Yugo Hayashi. 2020. Gaze awareness and metacognitive suggestions by a pedagogical conversational agent: an experimental investigation on interventions to support collaborative learning process and performance. *International Journal of Computer-Supported Collaborative Learning*, 15(4):469–498.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Daniel MK Lam. 2018. What counts as "responding"? contingency on previous speaker contribution as a feature of interactional competence. *Language Testing*, 35(3):377–401.

Holy Lovenia, Bryan Wilie, Willy Chung, Min Zeng, Samuel Cahyawijaya, Su Dan, and Pascale Fung. 2022. Clozer: Adaptable data augmentation for cloze-style reading comprehension. *arXiv preprint arXiv:2203.16027*.

Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, and Maxine Eskenazi. 2022. Interactive evaluation of dialog track at dstc9. *arXiv preprint arXiv:2207.14403*.

John Pill. 2016. Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing*, 33(2):175–193.

Juan-Pedro Rica-Peromingo. 2009. *The Status of English in Spain*, pages 168–174.

Carsten Roever and David W Dai. 2021. Reconceptualizing interactional competence for language testing. *Assessing speaking in context: Expanding the construct and its applications*, pages 23–49.

Carsten Roever and Naoki Ikeda. 2023. The relationship between l2 interactional competence and proficiency. *Applied Linguistics*, page amad053.

Emanuel A Schegloff. 1968. Sequencing in conversational openings 1. *American anthropologist*, 70(6):1075–1095.

Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings.

Isis H Settles, Martinque K Jones, NiCole T Buchanan, and Kristie Dotson. 2021. Epistemic exclusion: Scholar (ly) devaluation that marginalizes faculty of color. *Journal of Diversity in Higher Education*, 14(4):493.

Abdullayev Shaxobiddin. 2024. A discourse analysis of modal verbs in modern english: Patterns and functions. *Journal of new century innovations*, 50(2):145–147.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. *arXiv preprint arXiv:2005.00583*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Thomas Warren. 2017. *Cross-cultural Communication: Perspectives in theory and practice*. Routledge.

Steve Whittaker, Yvonne Rogers, Elena Petrovskaya, and Hongbin Zhuang. 2021. Designing personas for expressive robots: Personality in the new breed of moving, speaking, and colorful social home robots. *J. Hum.-Robot Interact.*, 10(1).

Jingxuan Wu and Carsten Roever. 2021. Proficiency and preference organization in second language mandarin chinese refusals. *The Modern Language Journal*, 105(4):897–918.

Foong Ha Yap and Anindita Sahoo. 2024. Versatile copulas and their stance-marking uses in conversational odia, an indo-aryan language. *Lingua*, 297:103641.

A Appendix

A.1 Software Availability

To contribute to the research community and facilitate further development and collaboration, we have made the source codes of our innovative annotation tool publicly available. The tool, designed with a focus on enhancing the efficiency and accuracy of data annotation processes, has been developed through meticulous research and development efforts. It incorporates a range of features tailored to meet the needs of researchers and practitioners working in fields that require precise and reliable annotation of datasets.

Accessing the Source Code

The source codes are hosted on GitHub, a platform widely recognized for its robust version control and collaborative features. Interested parties can access the repository at the following link: https://anonymous.4open.science/r/AnnotationTool2023-CFE1/README.md. This repository is intended for research usage, underlining our commitment to supporting academic and scientific endeavours.

Key Features and Capabilities

Our annotation tool stands out for its user-friendly interface, which simplifies the annotation process and allows users to work more efficiently. Among its key features are:

- **Customizable Annotation Labels:** Users can add their own set of labels to cater to the specific requirements of their projects.
- **Collaborative Annotation Support:** Facilitating teamwork, the tool allows multiple annotators to work on the same dataset simultaneously, ensuring consistency and reducing the time required for project completion.
- **Annotation History Tracking:** All the annotation history such as changes made can be tracked, and any further modifications can be done at any time for the user's convenience. Export Functionality: Annotated data can be exported in several formats, accommodating further analysis or use in machine learning models.

A.2 Pages View

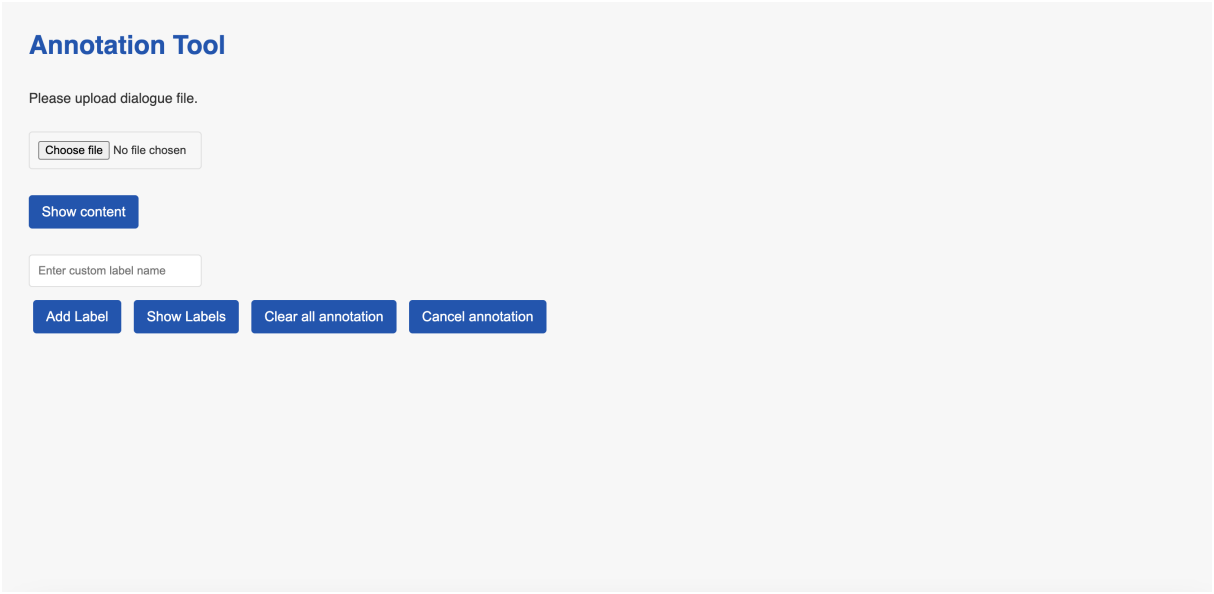


Figure 3: Annotation tool Demo

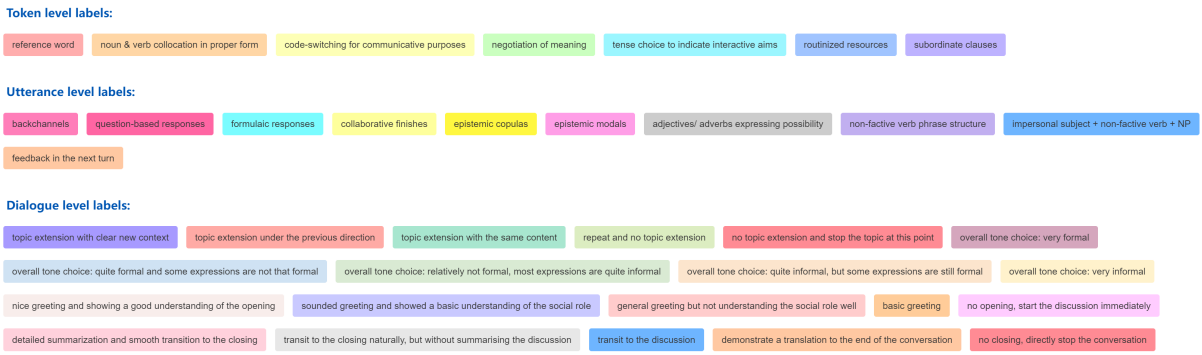


Figure 4: Hierarchical Label Assignment Demo

A.3 Manual

Annotation Manual for SLEDE Dataset

1. Introduction to the task

The research aims to investigate the interactive ability of second-language speakers of English through dialogue evaluation. The annotated data is about daily chat. You would be paired with another annotator with the same dialogue.

In your annotation, two types of dialogue tasks would be included in this study conducted by a pair-wise discussion by second language speaker participants. The first task is a storytelling task, in this part, two speakers will share some experience or what they want to deliver based on the instructions (e.g., *share some ideas on how you think of education in your life*). In the second task, two speakers need to solve a problem (e.g., *improve the experience of international students during their stay in Australia; and help to organize a welcome event*) through a joint discussion.

Dialogue of the two tasks were both transcribed into text and you are ready to annotate based on the text. **Videos will be provided if needed for correction of the text you are assigned.** Please notify the researcher, if you pick any misinformation in the transcriptions compared with the original recordings during your annotation.

2. Hierarchy sequence of the label

Label name	Label level	Label tag	example
reference word	Token level labels	[RA]	SPK_1 OK, that's all. SPK_2 That's all I think maybe we should switch from

			SPK_1 OK, wait for her [RA].
noun & verb collocation in proper form		[NVC]	SPK_1 No accidents. SPK_2 No accent. No, no. Like the the Beijing, Beijing accent. Yeah, that's that's the point. Yeah. So it's about the environment. SPK_1 Yeah, I think that's right [NVC].
code-switching for communicative purposes		[CS]	SPK_1 How do think of the educational policy in China? SPK_2 Hard to say, it depneds on different uh, **diqu (地区) [CS]** in China
negotiation of meaning (appropriate tense to show meaning)		[NM]	SPK_1 How you plan your next stage after graduate? SPK_2 I don't sure, maybe ask my presents whether they want to buy a house here or not.

			SPK_1 I' m going to see [NM] how my partner thinks.
tense choice to indicate interactive aims (politeness in talking/ social distance/ context variance) [TT]		[TT]	SPK_1 May [TT] I start first in this one? SPK_2 Ok.
routinized resources (projector construction)		[RR]	SPK_1 How you going today[RR]? SPK_2 Not bad.
subordinate clauses		[RC]	SPK_1 Everyone mandatory course. SPK_2 Yeah, yeah. It's a mandatory clause, so. So everyone needs to learn it. I think this is a pretty nice things to make [RC], make people like learn more things to have a big view for that.
backchannels	Utterance level labels	[BC]	SPK_2 That's all I think maybe we should switch from another park. SPK_1

			Oh [BC] SPK_2 Wait for her. We can do it myself. Let's see what the what is in that spoiler spoiler problem-solving discussion.
question-based responses		[QR]	SPK_2 Wait for her. We can do it myself. Let's see what the what is in that spoiler spoiler problem solving discussion. Instruction and at least. You need to with your partner and decide to. What solution to provide what kind of problem? Because she was solar problem together in this part. All we need to wait for right now, right? SPK_1 Yes, yes [QR]. Actually I need to pause here.
formulaic responses		[FR]	SPK_1 Good morning, I'm here to take in this task for Rena's study and SPK_2

			It’s nice to meet you here [FR] SPK_1 same
collaborative finishes		[CF]	SPK_2 No accent. No, no. Like the the Beijing, Beijing accent. Yeah, that’s that’s the point. Yeah. So it’s about environment. It’s. Yeah, I think that’s right. SPK_1 OK, that’s all.[CF]
epistemic copulas		[H1]	It seems [H1] to be a huge problem.
epistemic modals		[H2]	It might [H2] be a huge problem.
adjectives/ adverbs expressing possibility		[H3]	It is likely [H3] that this is a huge problem.
non-factive verb phrase structure		[H4]	This is possibly [H4] a huge problem.
impersonal subject + non-factive verb + NP		[H5]	These conclusions suggest a huge problem [H5].
feedback in the next turn		[FB]	SPK_1 A lot of people just don’t know. A second language. SPK_1 Ohh.[FB]

			SPK_2 Spanish. SPK_1 Yes.[FB]
topic extension with clear new context (change to utterance level, but more information context depends)	Dialogue level labels	[T1]	SPK_2 But you see that in China is all it was, like a lot of people just. SPK_1 Everyone mandatory course. SPK_2 Yeah, yeah. It's a mandatory course. So everyone needs to learn it. I think this is a pretty nice things to make, make people like learn more things to have a big view for that. And we can learn some beyond our own major studies in the uni. [T5]
topic extension under the previous direction		[T2]	SPK_2 But you see that in China is all it was, like a lot of people just. SPK_1 Everyone mandatory course. SPK_2 Yeah, yeah. It's a mandatory course. So ev

			eryone needs to learn it. I think this is a pretty nice things to make, make people like learn more things to have a big view for that. [T4]
topic extension with the same content		[T3]	SPK_2 But you see that in China is all it was, like a lot of people just. SPK_1 Everyone mandatory course. SPK_2 Yeah, yeah. It's a mandatory course. So everyone needs to learn it. [T3]
repeat and no topic extension		[T4]	SPK_2 But you see that in China is all it was, like a lot of people just. SPK_1 Everyone mandatory course. SPK_2 yeah. yeah. It's a mandatory course. [T4]
no topic extension and stop the topic at this point		[T5]	SPK_2 But you see that in China is all it was, like a lot of people just. SPK_1

			Everyone mandatory course. SPK_2 Yeah, yeah. [T5]
conversation opening		[CO1] [CO2] [CO3] [CO4] [CO5]	CO1: nice greeting and show a good understanding of conversation opening in social interactions. CO2: sounded greeting and show a basic understanding of the social role. CO3: general greeting and didn't demonstrate a good understanding of the social role. CO4: basic greeting. CO5: no opening just start the discussion immediately.
conversation closing		[CC1] [CC2] [CC3] [CC4] [CC5]	CC1: detailed summarization and smooth transition to the closing of the conversation. CC2: transit to the closing naturally, but without any summarization of the discussion. CC3: transit to of the discussion.

			CC4: demonstrate a translation to the end of the conversation. CC5: no closing, just stop the conversation.
overall tone choice: very formal		[OT1]	I’m very honoured to be here…
overall tone choice: quite formal and some expressions are not that formal		[OT2]	I’m more than happy to see you here today…
overall tone choice: relatively not formal, most expressions are quite informal		[OT3]	Happy to meet with you…
overall tone choice: quite informal, but some expressions are still formal		[OT4]	You know, meeting with you is quite happy…
overall tone choice: very informal		[OT5]	Hey, how’s going, nice today…

3. Label classifications and definitions

3.1 Token level

Label Category	Aspect	Definition
Reference word	Word choice	A reference word, also known as a referential word or referent, is a linguistic term used to describe a word or e

		xpression in a sentence that refers to or stands in place of something else in the text. Reference words are used to avoid repetition and to link different parts of a text together by indicating what a subsequent word or phrase relates to. Reference words can take various forms, including pronouns, demonstratives, and other words that replace or point to nouns or noun phrases.
Noun & verb collocation in proper form		Collocations are words or phrases that habitually occur together, forming a strong and natural linguistic association. In the case of noun-verb collocations, a particular noun is often paired with a particular verb due to convention, tradition, or linguistic patterns. These collocations contribute to the fluency, idiomaticity, and naturalness of language. Examples of noun-verb collocations: Make a decision: "I need to make a decision." Take a shower: "I usually take a shower in the morning." Catch a cold: "I hope I don't catch a cold." Give a speech: "She gave an inspiring speech."
Code-switching for communicative purposes		Code-switching for communicative purposes refers to the deliberate or subconscious a

		lternation between two or mo re languages or dialects wit hin a single conversation or utterance by bilingual or mu ltilingual speakers. This li nguistic phenomenon is emplo yed to fulfill specific comm unicative needs or function s, such as clarifying a poin t, expressing identity, sign aling solidarity or distinct ion, accommodating to the li stener's language preferenc e, or conveying concepts and emotions more effectively in one language over another. C ode-switching is not merely a random mixing of languages but a sophisticated communic ative strategy that reflects the speaker's linguistic com petence and cultural awarene ss, often used to navigate a nd negotiate the social and contextual dynamics of inter action.
Negotiation of meaning (appropriate tense to show meaning)	Contextual t ense usage	Negotiation of meaning refer s to the interactive process through which speakers of di fferent linguistic backgroun ds or competencies collabora tively work to understand ea ch other's intentions, messa ges, and linguistic expressi ons when communication break downs occur. This involves t he use of clarification requ ests, confirmation checks, c omprehension checks, and par aphrasing, among other commu nicative strategies, to ensu re mutual understanding is a chieved. The negotiation of

		meaning is a fundamental aspect of second language acquisition and communicative language teaching, highlighting the dynamic nature of language use and the active role learners play in constructing meaning through interaction.
Tense choice to indicate interactive aims (politeness / social distance/ context)		Tense choice to indicate interactive aims involves the strategic use of verb tenses by speakers to fulfill specific communicative goals or intentions within an interaction. This linguistic strategy encompasses the selection of present, past, future, or perfect tenses to convey nuances of time, mood, or aspect, directly influencing the interpretation and direction of the dialogue. Through careful tense selection, speakers can clarify the timing of events, express certainty or speculation about future occurrences, reflect on past experiences, or emphasize the continuity or completion of actions, all of which serve to enhance the clarity, persuasiveness, or relational dynamics of the communication. Tense choice, therefore, is not merely a grammatical decision but a deliberate tool employed by adept language users to navigate conversations and achieve specific interactive aims.

routinized resources (projector construction)	Interactional grammatical device	Routinized resources refer to patterns, practices, or tools that have become standardized and regularly employed within specific contexts or activities. These resources are often developed through repeated use over time, leading to a level of automation or routine in their application. In organizational or social settings, routinized resources help in streamlining processes, reducing the need for decision-making about routine tasks, and ensuring consistency in actions and outcomes. They can include documented procedures, established workflows, habitual practices, or even common language and scripts used in interpersonal interactions.
subordinate clauses		Subordinate clauses, also known as dependent clauses, are groups of words that contain a subject and a verb but do not express a complete thought and therefore cannot stand alone as a sentence. They function within a sentence by providing additional information to the main clause, to which they are connected by subordinating conjunctions (such as "because," "although," "when," "if") or relative pronouns (such as "who," "which," "that"). Subordinate clauses serve various roles in sentences, including acting as adjectives, adverbs, or nouns, and are essen

		tial for adding complexity, detail, and nuance to communication. Their use enables speakers and writers to articulate relationships of cause and effect, contrast, condition, time, and more, enriching the expressiveness and depth of language.

4. Questions to note

4.1 Q: What if I find multiple labels in one sentence/phrase/ token?

A: Label them all, and put all labels in the required formats indicated in this table.

4.2 Q: How to decide the tone in this dialogue?

A: After reading the whole dialogue, if you feel it is hard to decide based on your experience in daily communication, you can find the original video in the folder and watch it to find more information.

4.3 Do I need to correct the wrong points in the dialogue (e.g., grammatical error?)

A: No you don’t need to, if you find it hard to understand for a wrong point, you can refer to the original videos. Please keep the original content in the dialogue transcriptions.

5. Reference to consider when you start the annotate

5.1 http://compprag.christopherpotts.net/swda.html#tags

5.2 http://ling-blogs.bu.edu/lx390f16/classification/

5.3 https://aclanthology.org/D19-3021.pdf

A.4 Questionnaire

Your age

Your gender

- Male
- Female
- Non-binary / third gender
- Prefer not to tell

Current education/ job status

- highschool
- undergraduate
- graduate (Master)
- graduate research (PhD)
- employed

Your email address

Study abroad experience

Your home country:

- ○ China
- ○ Australia
- ○ [ ] Other country

How long have you stayed in Australia

- ○ less than 6 months
- ○ 6 months - 1 year
- ○ 1-2 year
- ○ 2-3 year
- ○ below 5 year
- ○ 5-10 year
- ○ over 10 year

Your first language

Besides your first language, what other languages have your learnt or can you speak?

Do you have any study abroad experience or stay abroad experience in English speaking countries?
If so, which country?

- ○ Yes
- ○ No

For how long have you spent your time as a study abroad student in English speaking country?

- ○ less than 1 month
- ○ 1-2 month
- ○ 3-6 month
- ○ 6-12 month
- ○ 1- 2 year (12-24 month)
- ○ 2-5 year (25-60 month)
- ○ more than 5 years (61 month)

What other countries have you stayed for the purpose of study/ work?

For how long?

- 1-3 month
- 3-12 month
- over a year (12 month)

Language proficiency

How easy is it for you to communicate in English? (English native speakers can ignore this question)

- Very easy, I can understand others and communicate in English **all the time**
- Mostly easy, I can use English well **in most cases**, but have trouble sometimes
- Sometimes easy, I can express myself and understand others **slightly more often** than not
- Sometimes difficult, I struggle to express myself and understand others slightly **more often than not**
- Mostly difficult, I struggle to express myself and understand others **most of the time** but occasionally I manage
- Very difficult, I struggle to express myself and understand others (nearly) **all the time**

How often do you use English

- ○ rarely
- ○ sometimes
- ○ often
- ○ always

How long do you use English in your communication everyday

- ○ less than 1 hour
- ○ 1 to 3 hour
- ○ 3 to 5 hour
- ○ above 5 hour

Block 4

Please score the below topics according to your preference

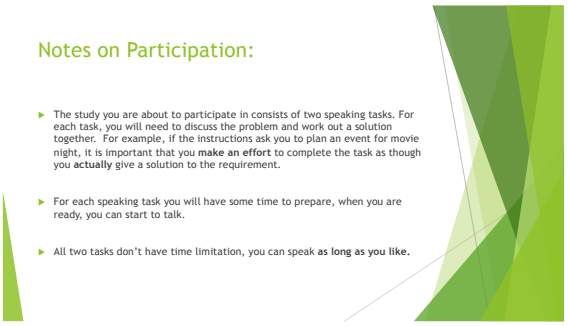
	0 1 2 3 4 5 6 7 8 9 10	Not Applicable
plan the schedule of an end-of-semester party	○	☐
improve the living experience for international students	○	☐

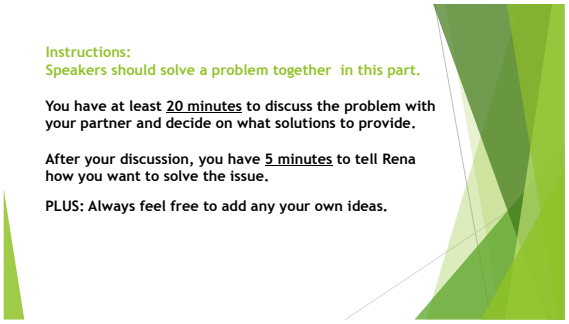
	0 1 2 3 4 5 6 7 8 9 10	Not Applicable	
decide a schedule for a two hour group discussion		☐	
select an elective subject in new semester		☐	
decide on how to distribute work for a formal presentation within group members		☐	
work out a solution for improving oral English in university communcations		☐	
give two solutions for improving the learning experience in BLS learning and teaching mode		☐	
plan a route for University of Melbourne Open-day tour for high shcool graduates		☐	

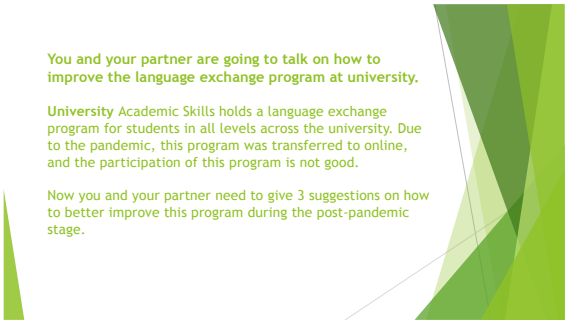
	0	1	2	3	4	5	6	7	8	9	10	Not Applicable
decide two subjects which you would recommed to newly commerenced students in your major	○											☐
provide two methods to help international students in adjusting local culture in Melbourne	○											☐

What are the common topics you focus or interested in daily chat?

A.5 Speaking Instruments







A.6 Experimental Result

Since our research results include a large number of figures and extensive data, we have organized them into a thorough document available on our GitHub repository. This helps us keep the information accurate and detailed for in-depth examination. To view all the results, the readers can visit this link: https://github.com/RenaGao/2024InteractiveMetrics/tree/main/2024ACLESLMainCodes_Results. Storing the results in this way makes them easy to navigate and ensures the quality and precision of the research are maintained.