Design and Development of a Human–Machine Dialog Corpus for the Automated Assessment of Conversational English Proficiency

Vikram Ramanarayanan

Educational Testing Service R&D
90 New Montgomery Street, Suite 1450, San Francisco, CA
vramanarayanan@ets.org

Abstract

This paper presents a carefully designed corpus of scored spoken conversations between English language learners and a dialog system to facilitate research and development of both human and machine scoring of dialog interactions. We collected speech, demographic and user experience data from non-native speakers of English who interacted with a virtual boss as part of a workplace pragmatics skill building application. Expert raters then scored the dialogs on a custom rubric encompassing 12 aspects of conversational proficiency as well as an overall holistic performance score. We analyze key corpus statistics and discuss the advantages of such a corpus for both human and machine scoring.

1. Introduction

We are seeing an increasing demand for conversational language learning and assessment solutions in today's educational marketplace. Conversational proficiency is a crucial skill for success in today's workplace [1, 2], and dialog-based language learning technologies are one solution capable of addressing and automating this need at scale [3]. However, such human–machine conversational technologies need to be able to provide useful and actionable feedback via either human or automated means to learners in order for them to be widely adopted. The former is labour-intensive and difficult to scale. The latter is more scalable, but requires a well-designed and annotated corpus in order to facilitate the training and refinement of automated scoring algorithms.

Let us contexualize prior work on corpus development for automated human–machine conversation scoring along three lines – availability, dialog task complexity and scoring detail/granularity – that in turn influence this paper's contributions. While there are many human-annotated corpora to enable automated scoring research for essays and short constructed text responses [4, 5, 6] and monolog speech [7, 8, 9, 10]) , there has been a relative dearth of corpora for the *interpretable* automated scoring of dialog. Evanini *et al.* (2015) examined a corpus of pseudo-dialogues, i.e., there were no branching dialog states; the system's response was fixed and did not vary based on the learner's response [11]. Other corpora involve more complex dialog tasks. Litman *et al.* developed a corpus of conversations where users were instructed to interact with a dialog system with the goal of finding laptops, restaurants or bus routes with certain characteristics, and human experts assigned these conversations a global CEFR scale score[1] [12]. Ramanarayanan et al. (2017) analyzed a corpus of human-machine interview data which was scored at the level of each dialog turn for three aspects of speech delivery – fluency, pronunciation and intonation [13]. While these are useful corpora, there are several granular aspects of the conversational proficiency construct[2] that such corpora do not explicitly score for, but are potentially crucial for providing more effective feedback to learners. Take for instance aspects pertaining to interaction – engagement, turn-taking and repair – which are a lot less well-studied as compared to others like delivery and language use. This paper presents a human-machine dialog corpus scored along 12 different sub-constructs of conversational proficiency (in addition to an overall holistic score) to bridge this gap and further research on human and automated scoring of language learners' conversational speech.

2. Data

2.1. Dialog Task

We considered a task-based dialog designed for language learners to practise and improve their conversational proficiency. This *Request Boss* task specifically focuses on making requests using pragmatically-appropriate language. The task instructs participants to imagine a scenario where they were going to interact with their (virtual) boss – Lisa Green, a representative image of whom is also presented on the screen – voiced by an interactive spoken dialog system. Their goals were to (i) schedule a meeting with her, and (ii) ask her to review presentation slides that they prepared earlier to discuss at the meeting. For more details, see [14].

2.2. Collection

We crowdsourced, using Amazon Mechanical Turk, the collection of 2288 spoken conversations of non-native speakers interacting with the *Request Boss* dialog appli-

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

[2] A *construct* in psychometrics can be defined as the knowledge, skills, and abilities that a given assessment is designed to evaluate or measure.

cation described above. To develop and deploy this application, we leveraged HALEF[3], an open-source modular cloud-based dialog system that is compatible with multiple W3C and open industry standards [3]. The HALEF dialog system logs speech data collected from participants to a data warehouse, which is then transcribed and scored.

2.3. User Experience Ratings

Following previous work [15, 16], to better understand how the system performs, we asked all non-native speakers to rate various aspects of their interactions with the dialog system on a scale from 1 to 5, 1 being least satisfactory and 5 being most satisfactory. We consider the following user experience metrics here:

1. *System latency*. This qualitative score measures perceived system response time. How debilitating is the average delay between the automated agent's response from the time the user finishes speaking to the conversation?
2. *System engagement*. A qualitative measure of caller's engagement with the task or the system, ranging from highly disengaged to highly engaged.
3. *System performance*. A qualitative measure of how the system performed as per speaker expectations and if system responses were appropriate.
4. *Perceived SLU*. A qualitative measure of the perceived spoken language understanding of the system, i.e., how well the system "understood" the non-native speaker.

3. Human Scoring

We had each of the $N = 2288$ dialogs scored by $R = 8$ human expert raters on a custom-designed scoring rubric. Our chosen experts all had significant experience in scoring various spoken and written assessments of English language proficiency. We based our rubric off a rubric for monolog speech scoring[4], and iteratively refined it to better capture construct characteristics specific to spoken dialog. The final conversational scoring rubric defined 12 sub-constructs under the 3 broad constructs of linguistic control, task fulfillment and interaction, apart from an overall holistic score. See Table 1 for more details.

We decided to triple score each dialog response in order to improve score reliability and minimize the effect of rater bias that might otherwise arise with a small number of raters. We chose our scoring design matrix (a N x R binary matrix where the $(n, r)^{th}$ entry is 1 if the nth dialog response is scored by the rth rater, and 0 otherwise) to satisfy the following criteria: (i) all raters had a commensurate number of responses to rate, and (ii) each rater scored a different subset of files. The second criterion introduces an element of randomization, which is important in order to prevent unwitting biases due to individual rater profiles creeping into the overall score analysis. We used a dynamic programming algorithm developed for exact counting and exact uniform sampling of matrices with specified row and column sums [17] in order to generate this scoring design matrix.

Table 1: *Human scoring rubric for interaction aspects of conversational proficiency. Scores are assigned on a Likert scale from 1-4 ranging from low to high proficiency. A score of 0 is assigned when there were issues with audio quality or system malfunction or off-topic or empty responses.*

Construct	Sub-construct	Description
Linguistic Control	Fluency	Examines to what extent the response includes pauses at appropriate locations to formulate ideas and good tempo with minimal hesitation.
	Pronunciation	Examines to what extent the response L1 influence and word-level pronunciation impacts intelligibility.
	Rhythm	Examines the extent to which appropriate sentence-level intonation and stress is used to convey meaning without hindering intelligibility.
	Grammar & Vocabulary	Examines the extent to which range of grammar structures and vocabulary is accurately used to express clear and precise meanings.
Topic Development	Topic	Examines to what extent the responses are uniformly on topic and relevant.
	Elaboration	Examines the extent to which arguments are developed taking into account dialog history and with minimal or no repetition.
	Structure	Evaluates the structure of the discourse and chain of reasoning, along with the appropriate use of discourse markers.
	Task	Evaluates how well the user accomplished the task over the course of the interaction.
Interaction	Engagement	Examines the extent to which the user engages with the dialog agent and responds in a thoughtful manner.
	Turn Taking	Examines the extent to which the user takes the floor at appropriate points in the conversation without noticeable interruptions or gaps.
	Repair	Examines the extent to which the user successfully initiates and completes a repair in case of a misunderstanding or error by the dialog agent.
	Appropriateness	Examines the extent to which the user reacts to the dialog agent in a pragmatically appropriate manner.
Overall Holistic Performance		Measures the overall performance.

4. Analyses and Observations

Figure 1 shows histograms of median score distributions for each sub-construct measured. We notice that a majority of our dataset comprises medium to high proficiency speakers, as seen from the larger percentage of 3 and 4 median scores assigned across sub-constructs. See Tables 2 and 3 for example dialogs that received an median holistic score of 4 and 2, respectively.

Table 4 shows inter-rater agreement statistics – Conger κ and Krippendorff α – for the human expert scores assigned to the data using the mReliability software tool[5].

[3] `http://halef.org`

[4] https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

[5] `https://github.com/jmgirard/mReliability`

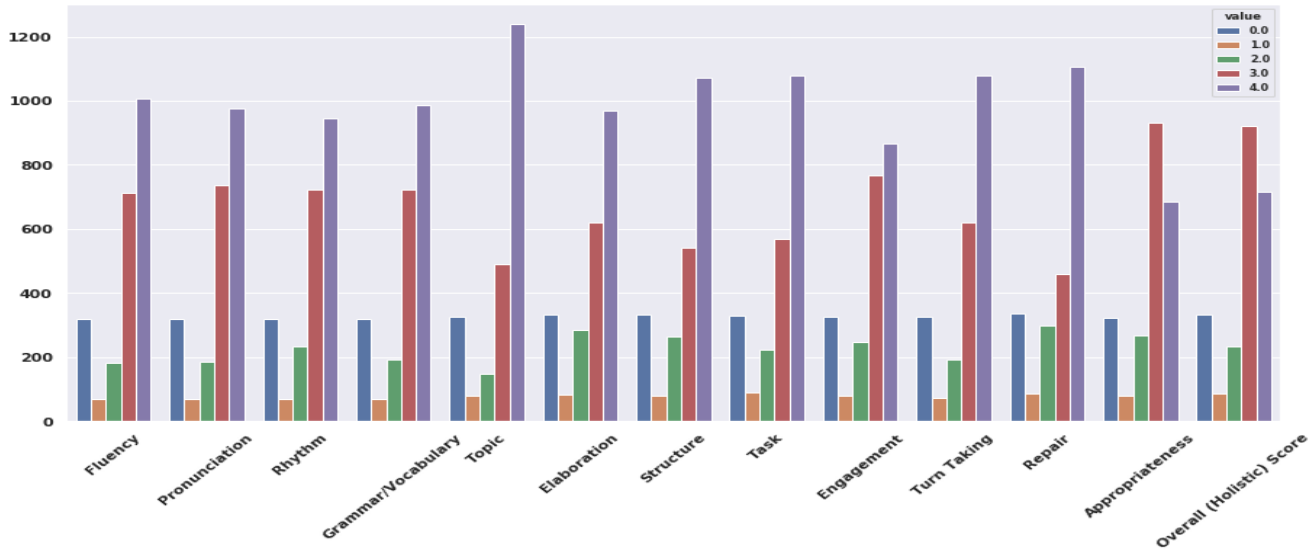

Figure 1: *Median score distributions computed over all raters.*

Table 2: *An example dialog that was assigned a median score of 4 across most dimensions (including overall score).*

Speaker	Dialog
System	Hello?
Human	Hi Lisa!
System	Hi! How's it going? What can I do for you?
Human	Uhm, I was wondering if you'd be able to have a meeting with me sometime this week and also if you could review my presentation slides before the meeting so that we can discuss them together?
System	Sure, no problem. I'll review the slides as soon as you send them to me and I'm available on Friday at 12 for a meeting. Does that work for you?
Human	Yeah, that works. Thanks!
System	Was there anything else you needed?
Human	No, thank you. That's everything.
System	Okay, see you on Friday then.

Table 3: *An example dialog that was assigned a median score of 2 across most dimensions (but 3 for all delivery sub-constructs).*

Speaker	Dialog
System	Hello?
Human	<background noise>
System	Hi! How's it going? What can I do for you?
Human	Yes I have meeting for next week. I also make a presentation.
System	Sure, I'm available on Friday at 12. Does that work for you?
Human	Yes, it's...it's okay for me.
System	Was there anything else you needed?
Human	No, I just...I just wanted to talk to you.
System	Sorry, I didn't get that. Was there anything else you needed?
Human	No this is what I want to talk to you. Thank you.
System	Sorry, I'm having trouble understanding you. I have to go, but please feel free to call back later.
Human	Okay thank you.

Recall that each dialog was scored by 3 out of 8 possible raters. We observe a moderate to high agreement between raters for all dimensions of the scoring rubric, which is unsurprising given that all our raters had significant experience in rating monologic speech data.

Figure 2 shows histograms of various user experience ratings as provided by each of our 2288 speakers who interacted with the dialog system. Speakers also tended to rate the system performance and understanding degree highly and self-reported a higher engagement rating. While they rated the system latency rating relatively lower than these other metrics on average, the histogram suggests that it was not debilitating to the conversations in general.

The scoring rubric laid out in Table 1 is comprehensive, but one might expect many of the dimensions to be correlated to each other to varying degrees. We therefore computed a Spearman correlation heat map to understand the extend of this correlation between dimensions of the rubric, as well as the user experience metrics described earlier. See Figure 3. We notice, unsurprisingly, that different sub-constructs associated with delivery are very highly correlated with each other, as are sub-constructs associated with topic development. Interaction sub-constructs are also highly correlated with aspects of delivery and topic development to varying degrees. *Repair*, appropriateness, *topic* and *task* related sub-constructs in particular have a lower Spearman correlation coeffienct with aspects of delivery, relative to other interaction and topic development sub-constructs, which suggests that these dimensions contain some information that is not present in the other channels. In addition, notice that all scores are generally uncorrelated with user experience metrics, suggesting that a speaker's proficiency did not depend on his/her user experience or perception of the system performance. This is important, because an absence of this would call into question the validity of the system as an effective interlocutor to measure a speaker's interactional competence. Further, the perceived latency and engagement metrics were not as highly correlated as the two system performance metrics, which suggest that these dimensions affected, but did not solely influence the overall performance of the dialog

Table 4: *Human Inter Rater Agreements for the same data expressed in Krippendorff α and Conger κ.*

Construct	Sub-construct	Human IRR κ	Human IRR α
Linguistic Control	Fluency	0.76	0.79
	Pronunciation	0.77	0.80
	Rhythm	0.76	0.78
	Grammar & Vocabulary	0.76	0.78
Topic Development	Topic	0.70	0.73
	Elaboration	0.76	0.75
	Structure	0.75	0.75
	Task	0.72	0.74
Interaction	Engagement	0.69	0.72
	Turn Taking	0.71	0.74
	Repair	0.73	0.72
	Appropriateness	0.70	0.72
Overall Holistic Performance		0.75	0.75

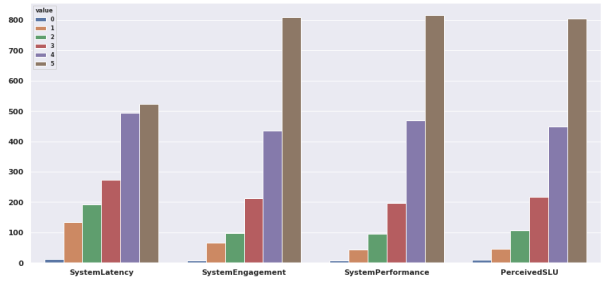

Figure 2: *Distributions of user experience ratings assigned by all speakers who interacted with our dialog system.*

system.

Figure 4 attempts to visualize the interdependence between each of the different sub-scores and the overall holistic score by non-linearly projecting the 13-dimensional vector of scores assigned to each of the 2288 dialogs into a lower dimensional manifold using the t-Stochastic Neighborhood Embedding (or t-SNE) technique [18]. The plot therefore projects each of the 2288 dialog samples from a point in 13-D space into 2-D space, with each dot in Figure 4 representing one dialog sample point. The color of the dots represent the different overall holistic score values assigned to the dialog represented by that dot. We observe that all the score classes are relatively well separated from each other with minimal overlap between them. We also note a bimodal distribution of points in score class 3, which suggests there are two distinct patterns of responses that produce a score level of 3.

5. Implications For Automated Scoring

This paper has presented a carefully-designed corpus of scored human–machine dialogs to facilitate automated scoring R&D for language assessment. Our analysis reveals that many proficiency subscores are correlated with each other, which will impact automated scoring algorithms, particularly the interpretability thereof. A set of features intended to capture *fluency* could also predict a score highly correlated with the trues scores for *rhythm* or *engagement* even if those features are not specifically designed to capture those aspects of the construct. This

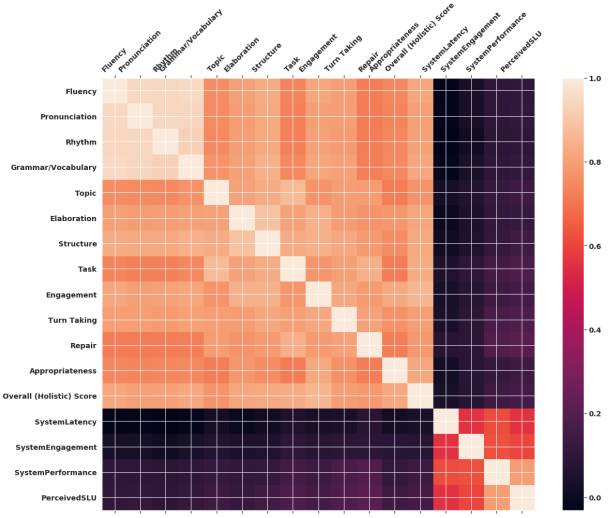

Figure 3: *Heatmap depicting the Spearman correlations between different ratings and user experience metrics.*

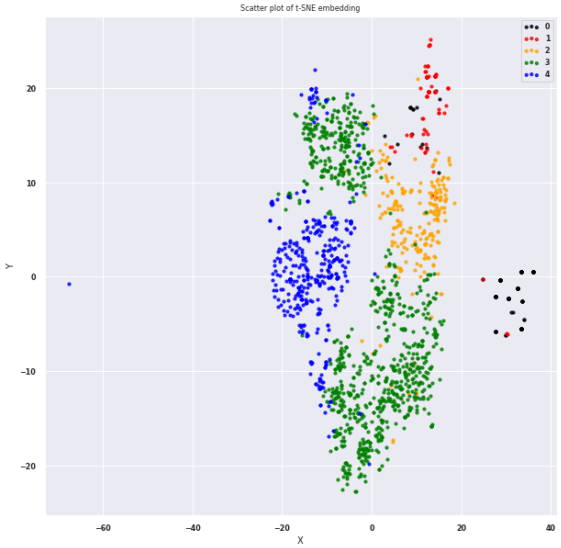

Figure 4: *Visualization of the t-SNE embedding manifold of all scores.*

could pose problems for interpretability, or in order to provide feedback to learners, particularly if the interpretation or feedback involves finding the salient features that were highly weighted in forming the predicted score.

The other point to consider is that certain dimensions are more abstract, and therefore harder to score for humans, as seen from the inter-rater agreement statistics. This is particularly the case for interaction-related constructs such as engagement, appropriateness and repair. This will in turn pose a greater challenge to automated scoring algorithms.

6. Acknowledgments

We thank Larry Davis and Veronika Laughlin for designing the initial version of the scoring rubric, and to Faye Weidner and Hillary Molloy for coordinating with 8 scoring experts to produce a refined version of both the rubric and final scores.

7. References

[1] T. G. Weldy and M. L. Icenogle, "A managerial perspective: Oral communi cation competency is most important for business students in the workplace jeanne d. maes," *The Journal of Business Communication*, vol. 34, no. 1, pp. 67–80, 1997.

[2] M. E. Oliveri and R. J. Tannenbaum, "Are we teaching and assessing the english skills needed to succeed in the global workplace?" *The Wiley Handbook of Global Workplace Learning*, pp. 343–354, 2019.

[3] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. V. Ivanov, Z. Yu, Y. Qian, and K. Evanini, "Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System," in *Multimodal Interaction with W3C Standards*. Springer, 2017, pp. 295–310.

[4] M. D. Shermis and J. Burstein, *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.

[5] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60–117, 2015.

[6] N. Madnani, A. Loukina, and A. Cahill, "A large scale quantitative exploration of modeling strategies for content scoring," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 2017, pp. 457–467.

[7] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech communication*, vol. 30, no. 2, pp. 83–93, 2000.

[8] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[9] X. Xi, D. Higgins, K. Zechner, and D. Williamson, "A comparison of two scoring methods for an automated speech scoring system," *Language Testing*, vol. 29, no. 3, pp. 371–394, 2012.

[10] S. Bhat and S.-Y. Yoon, "Automatic assessment of syntactic complexity for spontaneous speech scoring," *Speech Communication*, vol. 67, pp. 42–57, 2015.

[11] K. Evanini, S. Singh, A. Loukina, X. Wang, and C. M. Lee, "Content-based automated assessment of non-native spoken language proficiency in a simulated conversation," in *Proceedings of the Machine Learning for SLU & Interaction NIPS 2015 Workshop*, 2015.

[12] D. Litman, S. Young, M. Gales, K. Knill, K. Ottewell, R. van Dalen, and D. Vandyke, "Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of english," in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, p. 270.

[13] V. Ramanarayanan, P. L. Lange, K. Evanini, H. R. Molloy, and D. Suendermann-Oeft, "Human and automated scoring of fluency, pronunciation and intonation during human-machine spoken dialog interactions." in *INTERSPEECH*, 2017, pp. 1711–1715.

[14] V. Timpe-Laughlin, K. Evanini, A. Green, I. Blood, J. Dombi, and V. Ramanaranayan, "Designing interactive, automated dialogues for l2 pragmatics learning," in *Proceedings of the 21st workshop on the semantics and pragmatics of dialogue*, 2017, pp. 143–152.

[15] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, A. V. Ivanov, K. Evanini, Z. Yu, E. Tsuprun, and Y. Qian, "Bootstrapping development of a cloud-based spoken dialog system in the educational domain from scratch using crowdsourced data," *ETS Research Report Series*, 2016.

[16] V. Ramanarayanan, P. Lange, K. Evanini, H. Molloy, E. Tsuprun, Y. Qian, and D. Suendermann-Oeft, "Using vision and speech features for automated prediction of performance metrics in multimodal dialogs," *ETS Research Report Series*, 2017.

[17] J. W. Miller and M. T. Harrison, "Exact enumeration and sampling of matrices with specified margins," *arXiv preprint arXiv:1104.0323*, 2011.

[18] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.