Planning for a Corpus of Continuous Ratings of Spoken Dialog Quality

Nigel G. Ward

UTEP Computer Science Technical Report, UTEP-CS-19-28
March 20, 2019

Abstract

While many aspects of speech processing, including speech recognition and speech synthesis, have seen enormous advances over the past few years, advances in dialog have been more modest. This difference is largely attributable to the lack of resources that can support machine learning of dialog models and dialog phenomena. The research community accordingly needs a corpus of spoken dialogs with quality annotations every 100 milliseconds or so. We envisage a large and diverse collection: on the order of fifty hours of data, representing hundreds of speakers and many genres, with every instant labeled for interaction quality by one or more human judges. To make it maximally useful, its design will be a community effort.

This technical report is an edited version of a proposal to the National Science Foundation, submitted to the CISE Community Research Infrastructure Program in February 2019. I thank David DeVault, Milica Gasic, Kallirroi Georgila, Svetlana Stoyanchev, Tatsuya Kawahara, Olac Fuentes, and David Novick for helpful discussions.

1 Motivation: Enabling Faster Progress in Spoken Dialog Systems Research and Development

Dialog is a uniquely powerful form of interaction. Text-based interactions have their place, but if we want to get to know someone, negotiate plans, make lasting decisions, get considered advice, resolve a workplace issue, or have fun together, we usually seek a realtime spoken dialog, face to face or by phone.

Sometimes it would be helpful if computers could interact with us in the same way, for example as a coaches, tutors, or workplace assistants, but this is currently beyond the state of the art, and spoken language systems today are mostly confined to a few usage niches. Siri, for example, was initially designed to avoid dialog if at all possible, to instead get the job done with one response to one input, and this is still a common strategy. Indeed, at SLT 2018 (the IEEE Spoken Language Technology Workshop), one presenter, the lead developer of a well-known personal assistant system, revealed that for his system the average dialog length is 1.1 turns. While impressive in some ways, this is also testament to the difficulty of supporting true dialogs today.

In the research arena, researchers have shown how we can do better, producing prototype systems with amazing responsiveness. For example, Gratch produced a system capable of active listening better than most people, DeVault produced a system capable of in-game collaboration as fast as the average human, Litman demonstrated a system that could pick up on

subtle indications of a student's cognitive state and respond in ways that increased learning gains, Acosta produced a system that prosodically tailored its utterances to show empathy and thereby establish rapport, and Yu demonstrated how a robot through dialog could shape the user's attention [1, 2, 3, 4, 5]. These illustrate that **there are many ways in which the functionality and usability of dialog systems could be greatly improved.**

However such capabilities have not been taken up in commercial systems. One major reason is that their development involved custom corpora, careful policy design, and intense engineering and tuning. These do not scale. In contrast, the astounding recent advances in other areas of speech and language technology — speech recognition, speech synthesis, machine translation, speaker identification, emotion recognition, and many others — have been enabled by the application of deep learning models to large corpora. To enable rapid advances also in spoken dialog, the research community needs large corpora of data with suitable annotations to support deep learning approaches. In this project we propose to lay the groundwork for developing such corpora.

2 Research Landscape: How Spoken Dialog Systems are Built and Trained

Most task-oriented dialog systems are today designed by hand. This is because creating effective dialog today, even for something as familiar as automated banking, is as much an art as a science, both in the creation phase and in the refinement/tuning phase. Developers tend to be conservative and follow heuristics, such as avoid overlap at all costs, and avoid prosodic variation in favor of conveying everything explicitly with words. Such heuristics originated in the early days of dialog systems development, when the component technologies were far inferior to where they are today. They are ultimately rooted in informal quality judgments made by developers: they apply their intuitions about what users like and don't like, and may augment these with simple call-log statistics, for example to find reasons for glaring quality issues, such as those leading to dialog breakdown or call abandonment [6, 7].

In the research arena, by contrast, the current mainstream eschews questions of design in favor of developing ways to tune dialog systems from data. The general strategy is to devise cost functions and then create models that can be trained to optimize them. There are three main families of approaches.

1. For developing chat systems, standard practice is to train a system to match observed behavior, for example in choosing word sequences based on the degree of match to sequences seen in response to similar inputs in training data, such as chat corpora or movie screenplays. This approach is the core of neural conversational models, end-to-end dialog models and related approaches, which over the past three years have, from a few seminal papers [8, 9, 10, 11] grown to a tidal wave of work, with many hundreds of publications so far. While to date applied, as far as we know, only to text and to the content aspects of spoken dialog, it is only a matter of time before we see work generalizing these models to handle spoken interaction. At the same time, this family of approaches suffers from crippling problems with the metrics [12] (such as the commonly-used Bleu), which relate to simplistic concepts of match to observed behavior. In particular, human-human dialog interaction quality is not uniformly high, and there are many speakers and many actions which we do not want our dialog systems to mimic. (There are of course partial workarounds for this issue, including recording dialogs with an exemplary speaker, such as a champion-level customer service representative, and obtaining multiple response tracks, and then training the system to produce consensus behavior [13].)

2. For developing task-oriented dialog systems, there are more principled ways to optimize

policies to maximize value. As applied to information-seeking dialogs, such as restaurant recommendations, the original idea was that the quality of a dialog can be a simple function of the end result, such as whether the user gets a valid recommendation in a reasonable number of turns, and that policies can be optimized for this. This view can be elegantly formalized, and in combination with powerful learning techniques [14, 15, 16], for which open source toolkits exist [17], gives good results on benchmark problems of this type [18]. However, to date, this has been shown useful only for certain decision types, such as dialog act selection, for example, whether to confirm explicitly or implicitly. This technique is also data-hungry, requiring in practice the use of user simulations, which may not closely model actual human behavior. Such reinforcement learning models, while well matched to tasks where the semantic payload involves just one database lookup at the end, have also been extended to include intermediate reward signals, such as for the filling and confirmation of slots. Nevertheless, current models are not such a natural fit for dialogs involving richer or more dynamic semantics, as might occur, for example, in dialog in the service of collaborative realtime action, and it is not yet known how these models can be adapted to work for finer-grained decisions, such as those required for the types of responsiveness mentioned earlier.

3. While the above techniques are equally applicable to text or spoken dialogs, there is also work applying learning techniques to speech-specific phenomena. While judgments of overall dialog quality can give some insight into more specific decisions [19], most of those seeking to optimize local decisions in a principled way have sought to maximize a local quality metric. For example, for optimizing turn-taking, Raux defined the cost of gratuitous system silence as 1 per millisecond and that of a system interrupting the user as 5000 [20]. Most research on optimizing turn-taking based on data similarly uses arbitrarily-assigned costs [21, 22, 23]. Among other problems, these are not sensitive to the specific context — for example whether the interrupt was supportive, or prosodically designed to be tentative, or coming over a user vocalization that was prosodically designed as a post-yield particle — and have not been checked against human quality perceptions.

All of these approaches have lead to innovative, creative, and important work. But unfortunately research in machine learning approaches to dialog modeling has been limited by data availability, leading people to gravitate to optimizing things other than local quality, or using proxy data or weak dialog-level signals to do so. In contrast, we, like many in the community, think that **modeling and delivering high-quality, responsive interaction is high on the research agenda**, and we propose to face up to the challenge of developing data to support such research.

To put it simply, today we cannot build systems that directly optimize for high-quality behavior on short timescales. This is a major reason why we are failing to build systems that are pleasant to talk to. We could make a lot of progress on this if only we had a large quantity of the right kind of data; this would enable dialog systems to advance quickly like other fields of natural language processing.

3 Proposal: Continuous Annotations of Dialog Quality as a Resource for Training

We therefore want to develop a data resource with the following properties.

1. Explicit quality estimates. Rather than using indirect methods, if we know what is actually appropriate or inappropriate we can directly train systems to exhibit the desired behaviors.

2. Subjective quality estimates. Rather than knowing how some behavior matches a proxy metric of quality, if we know how real humans judges evaluate that behavior in that context, we can train more accurately.

3. Continuous quality estimates. Rather than just knowing the quality of the final dialog outcome, if we know throughout the dialog how well it is going, we can more directly train each decision. Comparison to human-human behavior is informative: we seldom just "rate your partner on a scale from 1 to 5 after the call" (as done for example in the Alexa Challenge), but instead continuously signal how pleased or displeased, frustrated, confused, and so on, we are at every moment.

Continuous rating is essential for two reasons. First, dialog systems, to be fully effective, need to make continuous choices. As they speak, they need to pay attention to small word choices and to the timing, pitch, intensity, durations and other attributes of their speech. Even when silent the system is (or should be) making choices, such as whether to continue to listen or bid to take a turn, or whether to wait to complete utterance planning or interpolate a filler. While currently most dialog systems are turn-based, there are no longer any fundamental latency constraints preventing fully incremental systems, and there is a good body of work on the value for users of doing so, as surveyed for example in [24, 25]. Continuous quality estimates will directly support such work. The second reason for continuous rating is that, even for standard turn-based, non-incremental dialog systems, continuous quality judgments will provide a more informative training signal, which will enable learning of better component-level performance and better dialog policies. In general, **this new resource will support the application of machine learning approaches to important problems** for which we currently have only partial, ad hoc, hand-crafted solutions.

Thus we propose to produce a resource with these three properties. Although there are dozens of corpora already available for building data-driven dialogue systems [26], there is no existing resource with continuous quality annotations; **this will be an entirely new kind of resource**.

For this we draw on two decades of work on dialog systems quality modeling. The first landmark was Walker's Paradise project at Bell Labs [27, 28], whose goals included enabling identification of the system properties that most impact bottom-line usability, and enabling automatic evaluation of system usability. The project elicited user satisfaction ratings at the granularity of dialogs, examined how these correlated with system properties, such as recognition error rate, and built predictive models.

Moller and colleagues at Deutche Telekom and the Technical University of Berlin built more accurate models of the relation between system properties, perceptual events, local quality judgments, and overall quality judgments [27, 29, 30, 31]. Ultes and colleagues [32, 33], at Ulm then Cambridge then Daimler Benz, extended the Paradise approach to include quality judgments at the granularity of "exchanges," that is, user-system response pairs. Their goals included supporting assessment of quality, of both human-human and human-computer dialogs, and automatic on-line adaptation of dialog strategy. They produced the "Lego" corpus of expert exchange-level annotations ("interaction quality judgments") over a subset of the CMU Let's Go corpus of callers seeking bus information from an automated system. Stoyanchev [34] used this framework and this dataset, together with proprietary data from two customer-service domains, and demonstrated good in-domain and fair cross-domain predictive performance. This indicates that quality judgments have much in common across diverse domains and genres, and thus the potential for quality judgments for a few genres to support the development of general models that are informative also for future genres. The EU SpeDial project is extending

this general line of work to consider also multimodal aspects. While basic research, such work is mostly directed at one application, that of determining, online, when the dialog quality is about to violate expectations and merits intervention by a human agent, either visibly or behind the scenes. Despite the promise, this line of work so far has been applied mostly to variants of the single-semantic-payload genre, such as dialogs seeking bus schedules or movie times.

Thus previous work helps us see what is needed and how to proceed. Data however has been lacking. This is a good point to stress the central role of suitably annotated data: researchers have found, again and again, that **adequate data is the major enabler of progress in spoken language research**. This is true also for dialog: data is typically a major cost and often the limiting factor in dialog research. The community is hungry for data, as seen by the fact that even the modestly-sized data sets of the Dialog State Tracking Challenges, even though they really support work on only one facet of dialog, have been highly influential in evoking and channeling research efforts, and in the impact of the Alexa Challenge (in which, however, Amazon doesn't provide the audio, only the ASR output with some timing information, so researchers participating in the competition can't use information from the speech signal such as prosodic features). Thus we expect a comprehensive, fully-open data set will be even more widely useful.

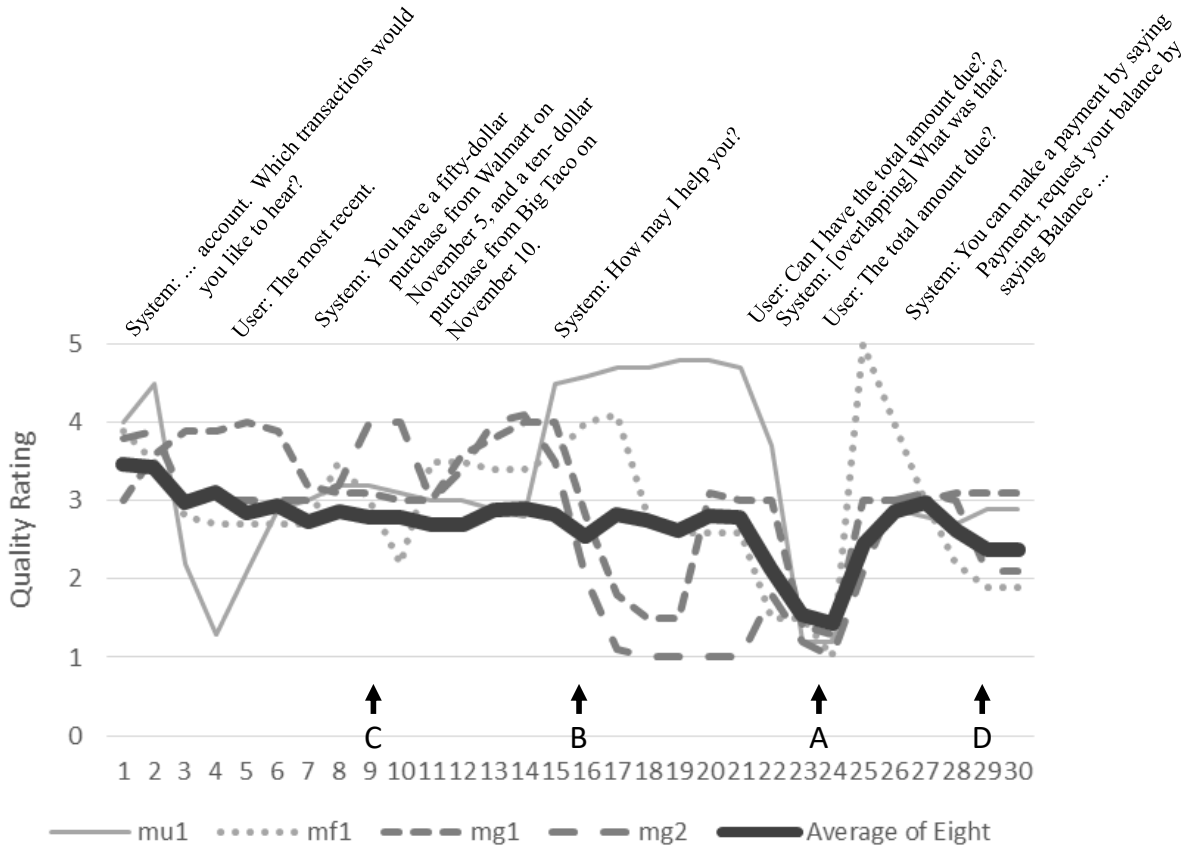
Beyond annotation of quality, many other aspects of dialog have been annotated. Most are semi-objective and focus on phenomena or taxonomies of linguistic interest, such as those in the NXT-format Switchboard Corpus [35]. Annotations of subjective properties include those for emotions, engagement, and rapport [36, 37, 38], generally done at the granularity of small slices or turns. To date there has been only one truly continuous annotation project, for emotions [39]. All these annotations were intended for use in training one or another specific component technology; in contrast, our aim is to support many uses.

At the same time, no set of annotated data will solve all problems: ultimately the final adaptation and tuning of dialog systems will always require domain-specific training data and/or ways to obtain live quality judgments with real-life users of those systems (the motivation behind, among other things, DialPort, DialCrowd, and similar projects [40, 41, 42]). However to support research in new functionalities and the training of better general models, there is nothing more valuable than adequately annotated, large-scale data sets. Thus we expect that the resources we will generate will enable the field to **finally realize the potential of learning-based approaches**, for spoken dialog, transforming the building of dialog systems from an art form to a solid technology based on optimizing behavior, with the result being highly efficient, widely available, pleasant-to-use dialog systems.

4 Sample Annotations and Uses

In this section we illustrate what might be learned from continuous quality ratings. In a research group meeting one day, with no warning, we had our colleagues and students rate a dialog fragment, including the 30 seconds seen in the figure.

The audio is available as 1:00-1:30 at www.cs.utep.edu/nigel/ccri.wav. This annotation was done without tools other than pencil and paper, without proper synchronization, and without instruction other than "judge quality", as our aim was only to discover issues and spark discussion. Even from this exceedingly crude first effort, however, interesting patterns emerge, as seen in the figure and learned from discussion.



In the figure, the salient drop at A was when the system interrupted the user. The divergence around B occurred because some raters liked the upbeat tone of voice, in contrast to the previous tedious tone during the information read-out, and others, paying attention to the dialog flow, felt that the bland replay of an earlier prompt seemed inappropriate for the dialog state, as evidenced by the user's subsequent long silence presumably reflecting confusion. The divergence around C reflected the appreciation some felt for the clear presentation of relevant information, and others' feeling that the pacing or prosody were awkward. The general slow drop starting around D occurred as the raters began to realize that the system had not understood the user and was backing off to a long listing of menu options.

Information like this could have many uses in training models or tuning dialog systems. Consider for example:

1) Turn-taking. Dialog systems are generally hardwired to avoid interruption. While many overlaps are acceptable — for example supportive overlaps, chiming in, backchanneling, and gently cutting off the user when understanding is hopeless — interruptions like the one at 1:21 are clearly not. With sufficient annotations like this, a system could learn to accurately model the costs of various kinds of interruption, and use that to properly tune its turn-taking policy.

2) Prompt Generation. Dialog systems designers generally operationalize maxims, like "provide sufficient information" and "be concise," and when these conflict, rely on their own intuitions or build a priori cost estimates into the language general algorithms. Continuous quality annotations will support development of more fine-grained objective functions, enabling more informed trade-offs.

3) Synthesis: Dialog systems developers often use prerecorded human voices to get suitable prosody, or else design the system's personality to make the restricted prosody of current

synthesizers less glaringly bad. The developers of synthesizers, in turn, generally optimize their systems for intelligibility and naturalness, with no intention (or ability) to tune them to be effective for dialog applications. With continuous quality annotations, dips in quality may be attributable to lapses in the quality of the synthesized voice, or its appropriateness in the context, enabling better training of synthesizers, in turn enabling the production of truly context-appropriate, goal-appropriate, and aesthetically-pleasing utterances.

4) Integrated Optimization. Dialog systems today are modular, with important decisions usually made in each module separately. However the effect on the user is not a linear sum of the effects of independent decisions. Consider for example, how the timing of a certain utterance, relative to the flow of turn-taking, may work well or be jarring depending on how it relates to the generated word sequence and to the synthesizer's prosodic choices. Continuous quality annotations will enable research that stradles conventional modularizations, supporting joint optimization, leading to improved consistency and better overall performance.

Thus **annotations like this could also support research of diverse types**, including, in addition to those mentioned above, accommodation/entrainment to the partner's dialog style, emotional coloring or responsiveness to help establish rapport, fine adjustments of timing and prosody to effectively establish roles and convey expectations, and entirely novel lines of inquiry.

5 Five-Year Goals

In five years we envisage a substantial corpus of spoken dialogs with quality annotations every 100 milliseconds or so.

Our vision is that in five years this corpus will be **supporting rapid advances in dialog models and systems**. We expect it to be in wide use by a large and diverse community of researchers: including both those already in the dialog field and newcomers with deep learning backgrounds who will bring new insights; including those seeking to develop better cost functions, better target functions, and better ways to prepare and select training data; including those doing both empirical descriptive studies of dialog phenomena and focusing on optimizing performance; including those working on improving components and those developing end-to-end systems; and including both academic and industry researchers.

The resulting flowering of research will lead to **new opportunities and approaches** for many challenging problems. These will include classic problems such as dialog act detection, next move selection, natural language generation, synthesizer control of prosodic parameters, and turn-taking. We also foresee advances on emerging topics involving adaptation, emotion, sentiment and stance-based responsiveness, timing to match the interlocutor's instantaneous cognitive load and receptivity, tracking the latent social state, and so on. This will not only enable further optimization of conservative dialog styles but also support explorations of new regions of the design space. The end result will be systems that are far more human-like, more effective, and more enjoyable to use.

In terms of support for CISE sub-disciplines, this will enable better dialog systems and other dialog-related language applications, such as speech-to-speech translation systems, and will also inform other work in speech and language technology, in computer-mediated communication, in computer-supported collaborative work, in assistive technologies, and in human-computer interaction more generally.

In terms of support for CISE research groups, domestically the academic and government sites most likely to use this resource in the short term include UCSC, USC-CS, USC-ICT,

UCD, MHC, Pitt, CMU, MIT, Columbia, UTEP, and ARL. On the industrial side we anticipate uptake by companies active in this space including giants such as Facebook, Uber, Amazon, Microsoft, Google, Apple, SRI, Nuance, IBM, Interactions, and Adobe, and smaller companies and start-ups such as Anticipant and b4.ai, to name just two. Adopters will include those aiming to build improved dialog systems and also those concerned with dialog abilities in the service of other tasks, such as human-robot interaction and other kinds of automation. As the barriers to entry in dialog systems research are shrinking rapidly, to the point where inexperienced student teams can create high-performing systems, as seen in the Alexa Challenge, we foresee **very broad uptake**.

6 Eighteen-Month Aims

In the planning phase, we need to 1) discover what is most needed and 2) what annotation practices are best, and 3) decide how to proceed. Our deliverables will be

- an initial small collection of continuous quality annotated dialogs
- a document on issues and alternatives in the collection of continuous quality annotations
- a document describing the data needs of the dialog research community
- an action plan and blueprint for a community resource, in the form of a full CRI proposal

The next two sections describe how we will produce these deliverables and meet these aims.

7 Community Involvement Plan

We will obtain **extensive community input, both broad and deep**. The end result will be a solid plan, wide community awareness of the possibilities enabled by this sort of data, and evidence of need that will enable a panel to recommend funding a full-scale effort.

1. We are organizing a special session, "Implications of Deep Learning for Dialog Modeling," at SIGdial 2019, to include position-paper talks and a panel discussion. In preparation we will invite all members of the SIG (the International Speech Communication Association's Special Interest Group on Discourse and Dialog) to provide input, including a prompt for comments on a resource like this. At the panel we will hand out detailed feedback forms, and afterwards follow up individually with people with specific needs, suggestions, or concerns.

2. We plan to run focus groups of 3-5 people, to discuss needs and possibilities in this space. We will do this widely, as we aim for diverse impact, supporting research not only on core dialog issues, but also for at improving the utility of component technologies for dialog systems needs. Advisory Board members and additional consultants will be crucial here, running groups at most or all of: Interspeech, the Speech Synthesis Workshop, the IEEE Automatic Speech Recognition and Understanding Workshop, the Natural Language Generation Conference, the Affective Computing and Intelligent Interaction Conference, Speech Prosody, the International Conference on Multimodal Interaction, Computational Linguistics and Clinical Psychology, Human-Agent Interaction, and Human-Robot Interaction.

3. We will complement these face-to-face interactions with broader efforts to understand the "market" for our corpus. We intend to start with a mailing, not duplicating that for the SigDial panel, inviting participation in a survey. The survey would ask a few simple questions (such as: What are your needs for data for training? What does the field need overall? If we built such a corpus how likely would you be to use it? What would you use it for?

What factors would be most important in your decision of whether to use it?) and contain an open-comment field. We will then follow up individually with those who have important comments or novel perspectives, or who wish to get involved. However since most people will not respond to a generic mail or survey, we will individually contact selected researchers, selected for diversity (of area, of seniority, of research style, of affiliation, etc.), asking them specific questions tailored to their expertise and what they can help us understand about the needs and the possibilities. Since respondents may be concerned that revealing information about data needs might also reveal future research intentions and commercial plans, we will do this sensitively, with discussions by phone/Skype rather than email whenever possible. These conversations will be wrapped up by December.

4. Overlapping these efforts, we will prepare and release a small set of continuous quality annotations, tentatively by January 2020. This will enable diverse research groups to explore, experiment, and give focused feedback. To encourage them to do so, we may organize a special session on continuous annotation approaches or the uses of such annotations at Interspeech 2020 or another venue. We may also set up and run one or two challenge tasks involving the pilot dataset. One obvious challenge task would be to create a system best able to automatically predict quality judgments from raw data [34], and not only within one genre but also across genres. We will however take care not to give the impression that this resource is intended to be used for only one purpose.

5. We will host a small two-day hands-on workshop, tentatively in El Paso in December. This is because we need participants' focused attention to work out the implications of low-level choices in annotation procedures for utility for various tasks. There will be time for 1) exploratory annotation by the participants of data in several genres, 2) discussion of the differences observed in the annotations, as a function of methods, tools, instructions and annotator individual differences, 3) whiteboard-level design of machine learning approaches to both end-to-end and component-level optimization for dialog systems using such data, 4) blue sky discussions of the future of the field, leading to 5) a high-level description of what the full corpus should look like.

6. We will produce a strawman blueprint for the final corpus and circulate it to already-engaged individuals for comments in February 2020.

7. Based on discussion we will create a serious blueprint proposal in July 2020 and widely circulate it. This will document decisions regarding: which specific corpora to annotate (considering issues like availability without licensing restrictions, existence of transcriptions or other complementary annotations, relevance to touchstone tasks, relevance to emerging dialog use cases, and diversity: of interactants (ages, genders, personalities, native languages etc.), of overall quality levels (spanning the spectrum from frequently highly problematic to master communicator), and genre (including at least task-oriented dialogs, chat dialogs, and collaborative dialogs) and so on); whether to include a "core" of tens of hours of data in the same genre, recorded under the same conditions, annotated by one consistent annotator; whether to organize the main project using a distributed, federated, or franchised model, or whether to do almost everything at one site; whether to offer a service component, in which we provide annotations for (almost) any proffered corpus, in exchange for rights to release the corpus and annotation; the distribution mechanism; and so on.

8. We will host a forum on this blueprint, at or co-located with a relevant conference or workshop. Our aim will be uncover any new considerations or reasons to adjust priorities, methods, and plans, and to inform the community. We know that there are many ways to look at dialog and many approaches to training, so we do not expect perfect consensus, but

we will strive to be inclusive and make our data collection as useful as possible to as many groups as possible. **We will measure success** for the planning phase overall by the number of people who attend or otherwise indicate the intention to use the resources we will create: high attendance will indicate that we have successfully identified real, widespread, important needs and a widely-shared feeling that continuous quality annotation can help.

We will schedule these activities so as to be ready with a full CCRI proposal for the January 2021 deadline. However it is possible that by December 2019 we will already have a broad consensus and see no roadblocks, and in that case we may submit a proposal in the January 2020 round. Either way, we will perform the community engagement during this planning phase so as to set a good foundation for outreach during the main project.

8 Planned Explorations in Annotation Methods

To complement and support our community-engagement work, we will do some in-house work.

Adopting an agile strategy, we will produce a few pilot annotations right away. Such annotations can be gathered efficiently by using physical input devices such as dials, joysticks, or sliders [39]. We will organize and release what we gather from the start, as any resources are better than no resources, and we will not delay sharing for lack of the Holy Grail of perfect methods. We will improve as we go along, based on feedback from our annotators, our own observations of the process, and numerous small statistical studies.

The initial small collection, released to the community, will serve to facilitate informed discussions. The details will be determined later, but at the moment we foresee the need to include both human-human dialogs and human-computer dialogs, since they have different failure modes and success modes. UTEP has small collections of both types, in diverse types, which have no licensing issues, which may be adequate for this phase [29, 43, 44, 45, 3].

In this planning phase, we will address any important issues that arise. Currently we see foresee the need to work on at least four families of issues:

a) Issues of Agreement and Individual Differences

Previous work indicates that interaction quality can be evaluated with good levels of agreement among "expert" raters [32], but we are also aware of unpublished work showing that agreement with and among novice raters, including participants (system users themselves, that is to say first-party raters) are much lower, in some cases with inter-annotator correlations as low as 0.2. While there are many techniques and tricks one can use to reduce variation and boost agreement [46], here we plan to explore and use only well-justified methods, for three reasons. First, quality ratings are opinions, and we cannot expect people to always agree. This also implies that we will not be able to validate our procedures in any rigorous way. Second, the annotations are intended for use as input to powerful learning models, which will trivially be able to compensate for differences in scaling, response lags, and other factors. Third, individual differences in judgments can be a valuable source of information. Our preliminary studies suggest that some people are not sensitive to aspects that others consider to be important. Including some annotations from such individuals will enable learning to distinguish bad-for-everyone behaviors from those that are irritating just for some. Over the long term, we can foresee that diversely annotated data could be useful for identifying user types or dimensions of user variation, enabling systems to adapt their behavior to match a derived model of the quality preferences of a specific new user.

In the planning phase we will, without aiming to resolve all such issues, do several explorations into quality judgment differences. We will explore how well annotators from different

perspectives agree and where they tend to differ and why. For example, we will explore how judgments differ as a function of levels of experience and expectation, as frequent users of dialog systems will probably systematically differ in their judgments from those with less experience. We will especially focus on differences in judgments by first-person annotators (retrospectively examining their own behavior in an interaction in which they had earlier participated), by second-party annotators (retrospectively examining the behavior of a dialog partner), and the (far more affordable) third-party annotations.

b) Quality Dimension Issues

Quality is a complex construct. In this project it is not a primary goal to improve the scientific understanding of the nature of quality judgments. We know that objective notions of quality — as in the ISO 9000 definition, "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" — do not directly map to subjective quality judgments. We also know that perception of quality is not direct but mediated by comparison of observed behavior to expected behavior [30], that judgments of quality involve multiple dimensions, and that ultimately quality and value depend on the intended purpose of the interaction or system. Quality perceptions also are highly dependent on the perspective taken, as the goals of service provider, system developer and user may diverge. Further, a dialog behavior at a specific time may seem poorly chosen in the local context, but may have strategic value for the overall dialog goal.

Previous work indicates that interaction quality, our focus here, can be evaluated independently of quality of content [32]. Our working hypothesis is that a crude overall measure based on fairly quick judgments by fairly naive annotators will be useful for many practical purposes, and that asking annotators to judge "quality" without explanation will usefully result in overall judgments, rather than those focused on one or another specific aspect or specific personal preferences.

In the planning phase we will do some small experiments to determine whether to annotate, for at least some of the data, not just a single dimension of quality but also more specific dimensions — such as efficiency, naturalness of flow, pragmatic appropriateness, politeness, or effectiveness. We expect them all to correlate to some degree, but to also find that some are distinct enough to annotate for separately, to provide added value for purposes of optimizing some system components or decision types.

c) Issues of Scope and Scale

In data collection for machine learning, bigger is always better, but we do not wish to spend federal dollars for data increments of only marginal value. Accordingly we need to carefully estimate how much work to propose for the main project. One consideration is that some problems in dialog can be solved, for a specific domain, by deep learning with only 11 hours of training data [23] (but see also the caveats in [47]). Another consideration is that deep learning work generally requires millions of data points to outperform designed models, which for dialog, assuming a data sample every 100 milliseconds, means 60+ hours of annotation. If these estimates prove correct, then the main project may be quite short in term and quite affordable for the taxpayer.

In the planning phase we will explore the cost-benefit ratio of collecting more data versus using existing data for bootstrapping (automatic generation of annotations using new data), either fully automatically or in an active learning mode. We will also work to document our methods as "standards" so that others in future may annotate consistently with our practice, relieving us of the need to be permanently available for annotation in the out years.

d) Practical and Community Issues

There are many issues which cannot be answered by technical means alone, but whose resolution, in consultation with the community, can be informed by in-lab exploratory work. These include: the choice of hardware and software infrastructure for collecting the annotations; the decisions of what fraction of multimodal and/or situated data, what fraction of human-human data, and what fraction of non-English data to include; the decision of for what fraction of the data the annotators should be given transcripts or video recordings, rather than the audio alone; and the decision of what fraction of the data annotations should be based on first impressions, versus listening once and then annotating during relistening, versus annotating while replaying at will.

These explorations will result in, among other things, a technical report on the processes and properties of continuous quality annotation, to serve as a resource for others in the future considering similar work. They will also enable us identify the issues with and limitations of continuous quality annotations obtained in this way, likely to be covered as part of the same report. Most pertinently, these explorations will enable us, with the assistance of the Advisory Board, to decide how to proceed. Thus **we already have identified the likely main issues and have a plan** for how to resolve them, not definitively, but to the point of being able to propose a well-defined full project.

9 Broader Impacts

We envisage a future where dialog systems are easy and pleasant to interact with, and where spoken dialog, with its unique advantages for establishing rapport, motivating, and coordinating action, is widely used. Our work will support the development of new applications in which people and computer systems work together in close collaboration, smoothly and efficiently, leading to new capabilities and accomplishments [48, 24]. While these advantages are not critical for all applications, and we certainly do not advocate fussing over microbehaviors for their own sake, it is often critical to get the little things right in order to show attentiveness, effectively convey information, be accepted, and generally leverage human knowledge of dialog norms and their well-learned interaction skills [49].

As autonomous systems become more capable, proliferate more widely, and interact more closely with humans, this will be of increasing importance. Such abilities will be even more useful in future applications when the interaction is social in nature or when the task to be performed is ill-defined, poorly understood, or too complex to resolve in a single input-output exchange. Tutoring systems, self-driving cars, kitchen robots, mobile robots, companion systems, entertainment agents, and so on can all be made more competent, effective, and acceptable. For example, consider coaching-style tutoring for grade-school math, where it has been suggested that the benefit of tutoring lies in the way that "the *highly interactive* nature of tutoring *itself* promotes learning" [50] (emphasis added). Consider also future caregiving robots, where users are very unlikely to trust a robot near their body unless it has at least the minimal interaction skills that indicate basic situation awareness and social competence [24]. Consider in addition assistive and augmentative communication systems, enabling individuals to be more effective and have more rewarding social interactions. Dialog resources with continuous quality annotations will in the long term enable the development of such abilities and systems.

We also expect our work to support improved dialog systems for conventional applications over the near term.

Beyond interactive systems, we expect these results to also support other applications that

involve dialog: simultaneous interpretation, automatic social role detection, speaker recognition, dialog outcomes prediction, language proficiency evaluation, clinical diagnosis, and many more. In some cases dialog behavior styles are informative, and by better modeling them, our work will support automatic detection of deviations from the typical patterns, which will in turn support these tasks. For other tasks the dialog factors are chaff obscuring the more relevant social signals, and a better understanding will enable these to be factored out.

It is sometimes thought that speech technology is a solved problem and thus that government involvement is no longer needed. The vast resources available to the tech giants are often mentioned in support of this view. However the tech giants themselves do not seem to think this way, and indeed, two of the biggest are explicitly seeking fresh ideas for how to escape from our current local minimum in dialog technology, in the Alexa Skills Challenge and the Microsoft Dialog Challenge, respectively. The new resources we will provide will **support the wider community** in the development of new approaches to classic problems and will open the door to creative work on new challenges.

We also note that while, to date, the commercialization of spoken dialog systems has been an American success story, there is a great deal of important research also in Japan and Europe, and developing a shared resource that is primarily in English may help keep our nation at the top, and thus able to fully reap the social and economic benefits.

10 Scientific Importance

This project will also more broadly advance the science of human interaction.

There is great interest in psychology and other social sciences in how realtime interpersonal coordination is accomplished, and there have been many qualitative and quantitative studies of the processes involved in the dynamic interplay between interactants at the sub-second level. However the field still lacks detailed models [51, 52, 53, 54, 55, 56, 57, 58, 49, 59, 60, 61]. Spoken dialog is of central importance here [62], as very likely the simplest example of prediction and coordination in social interaction, and thus a good area in which to study these topics. This corpus may help enable the enterprise of modeling real-time interactive behavior to rise to a new level of specificity.

Improving the scientific understanding of interaction will eventually help individuals and groups overcome obstacles to communication, helping them participate more fully in society. There are many practitioners applying micro-analysis and related methods for couples therapy, data mining for effective behaviors [63, 38], optimizing service interactions, improving workplace communication, and other purposes. However not all of the advice given is truly evidence-based, a problem that a corpus of dialog quality judgments will help overcome. There is also a great public appetite for knowledge of how to communicate more effectively, with many practical questions as yet unanswered. Discovery of valued dialog patterns, as enabled by our corpora, will help members of the public improve their dialog behavior, either directly or thanks to improved coaching. Many non-native speakers wish to learn how to be more polite, more effective, or more accepted. Many native speakers would like to learn to perform better in challenging roles, such as selling, flirting, persuading, counseling and interviewing. Appropriate dialog skills are important for social success, and deviations from norms can lead to unhappy outcomes [64, 65, 66]. Better understanding of dialog behaviors and how they are perceived may also help us better understand the nature of impairments, as in autism [67], and ultimately help us to improve the specificity of diagnoses and to design better treatments.

References

[1] J. Gratch, N. Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, R. van der Werf, and L.-P. Morency, "Can virtual humans be more engaging than real ones?," *Lecture Notes in Computer Science*, vol. 4552, pp. 286–297, 2007.

[2] M. Paetzel, R. Manuvinakurike, and D. DeVault, "So, which one is it? the effect of alternative incremental architectures in a high-performance game-playing agent," in *Sigdial*, 2015.

[3] J. C. Acosta and N. G. Ward, "Achieving rapport with turn-by-turn, user-responsive emotional coloring," *Speech Communication*, vol. 53, pp. 1137–1148, 2011.

[4] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, pp. 1115–1136, 2011.

[5] Z. Yu, D. Bohus, and E. Horvitz, "Incremental coordination: Attention-centric speech production in a physically situated conversational agent," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 402–406, 2015.

[6] T. Paek, "Toward evaluation that leads to best practices: Reconciling dialog evaluation and resarch and industry," in *Proceedings of the ACL Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pp. 40–47, 2007.

[7] C. J. Hayes, M. Moosaei, and L. D. Riek, "Exploring implicit human responses to robot mistakes in a learning from demonstration task," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pp. 246–252, IEEE, 2016.

[8] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[9] J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, and J. Weston, "Evaluating prerequisite qualities for learning end-to-end dialog systems," in *ICRL*, 2016.

[10] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *AAAI*, vol. 16, pp. 3776–3784, 2016.

[11] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[12] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[13] L. Huang, L.-P. Morency, and J. Gratch, "Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior," in *9th Int'l Conf. on Autonomous Agents and Multi-Agent Systems*, 2010.

[14] S. Singh, D. Litman, M. Kearns, and M. Walker, "Optimizing dialog management with reinforcement learning: Experiments with the NJFun system," *Journal of Artificial Intelligence Research*, vol. 16, pp. 105–133, 2002.

[15] S. Young, "Using POMDPs for dialog management," in *IEEE/ACL Workshop on Spoken Language Technology (SLT 2006)*, 2006.

[16] S. Young, M. Gasic, B. Thomson, and J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," *Proceedings of the IEEE*, vol. 101, pp. 1160–1179, 2013.

[17] S. Ultes, L. M. R. Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gasic, *et al.*, "Pydial: A multi-domain statistical dialogue system toolkit," *Proceedings of ACL 2017, System Demonstrations*, pp. 73–78, 2017.

[18] J. D. Williams, M. Henderson, A. Raux, B. Thomson, A. Black, and D. Ramachadran, "The dialog state tracking challenge series," *AI Magazine*, vol. 35, no. Winter, pp. 121–123, 2014.

[19] S. A. Chowdhury, E. A. Stepanov, and G. Riccardi, "Predicting user satisfaction from turn-taking in spoken conversations," in *Interspeech*, pp. 2910–2914, 2016.

[20] A. Raux and M. Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in *NAACL HLT*, 2009.

[21] N. G. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, pp. 1177–1207, 2000.

[22] I. de Kok and D. Heylen, "A survey on evaluation metrics for backchannel prediction models," in *Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

[23] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *Sigdial*, 2017.

[24] N. G. Ward and D. DeVault, "Challenges in building highly interactive dialog systems," *AI Magazine*, vol. 37, no. 4, pp. 7–18, 2016.

[25] V. Tsai, T. Baumann, F. Pecune, and J. Casell, "Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants," in *Proceedings of the 2018 International Workshop on Spoken Dialog System Technology*, 2018.

[26] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems: The journal version," *Dialogue & Discourse*, vol. 9, no. 1, pp. 1–49, 2018.

[27] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "Evaluating spoken dialog agents with paradise: Two case studies," *Computer Speech and Language*, vol. 12, pp. 317–348, 1998.

[28] M. Walker, C. Kamm, and D. Litman, "Towards developing general models of usability with Paradise," *Natural Language Engineering*, vol. 6, pp. 363–377, 2000.

[29] N. G. Ward, A. G. Rivera, K. Ward, and D. G. Novick, "Root causes of lost time and user stress in a simple dialog system," in *Interspeech*, 2005.

[30] S. Möller and N. Ward, "A framework for model-based evaluation of spoken dialog systems," in *Sigdial*, 2008.

[31] S. Möller, K.-P. Engelbrecht, and R. Schleicher, "Predicting the quality and usability of spoken dialog services," *Speech Communication*, vol. 50, pp. 730–744, 2008.

[32] A. Schmitt and S. Ultes, "Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction," *Speech Communication*, 2015.

[33] S. Ultes, P. Budzianowski, I. Casanueva, N. Mrkšic, L. Rojas-Barahona, P.-H. Su, T.-H. Wen, M. Gašic, and S. Young, "Domain-independent user satisfaction reward estimation for dialogue policy learning," in *Proceedings of Interspeech*, pp. 1721–1725, 2017.

[34] S. Stoyanchev, S. Maiti, and S. Bangalore, "Predicting interaction quality in customer service dialogs," in *Advanced Social Interaction with Agents: Proceedings of the 8th International Workshop on Spoken Dialog Systems* (M. Eskenazi, L. Devillers, and J. Mariani, eds.), pp. 149–159, Springer, 2019.

[35] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, *et al.*, "The NXT-format Switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language Resources and Evaluation*, vol. 44, pp. 387–419, 2010.

[36] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[37] S. Al Moubayed and J. Lehman, "Toward better understanding of engagement in multiparty spoken interaction with children," in *Proceedings of the International Conference on Multimodal Interaction*, pp. 211–218, 2015.

[38] M. A. Madaio, R. Lasko, J. Cassell, and A. Ogan, "Using temporal association rule mining to predict dyadic rapport in peer tutoring.," in *Educational Data Mining*, 2017.

[39] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[40] T. Zhao, K. Lee, and M. Eskenazi, "Dialport: Connecting the spoken dialog research community to real user data," in *Spoken Language Technology Workshop (SLT), IEEE*, pp. 83–90, 2016.

[41] K. Lee, T. Zhao, A. W. Black, and M. Eskenazi, "Dialcrowd: A toolkit for easy dialog system assessment," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 245–248, 2018.

[42] D. Schlangen, T. Diekmann, N. Ilinykh, and S. Zarrieß, "slurk–a lightweight interaction server for dialogue experiments and data collection," in *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue (AixDial/semdial 2018)*, 2018.

[43] N. G. Ward and S. D. Werner, "Data collection for the Similar Segments in Social Speech task," Tech. Rep. UTEP-CS-13-58, University of Texas at El Paso, 2013.

[44] N. G. Ward and P. Gallardo, "A corpus for investigating English-language learners' dialog behaviors," Tech. Rep. UTEP-CS-15-33, University of Texas at El Paso, Department of Computer Science, 2015.

[45] N. G. Ward and S. Abu, "Action-coordinating prosody," in *Speech Prosody*, 2016.

[46] A. Spirina, O. Vaskovskaia, T. Karaseva, A. Skorokhod, I. Polonskaia, and M. Sidorov, "Analysis of interaction parameter levels in interaction quality modelling for human-human conversation," in *International Conference on Speech and Computer*, pp. 130–140, Springer, 2017.

[47] N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, "Turn-taking predictions across languages and genres using an LSTM recurrent neural network," in *IEEE Spoken Language Technology Conference*, 2018.

[48] D. Bohus, E. Kamar, and E. Horvitz, "Towards situated collaboration," in *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pp. 13–14, Association for Computational Linguistics, 2012.

[49] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, 2007.

[50] S. Katz, P. Albacete, M. J. Ford, P. Jordan, M. Lipschultz, D. Litman, S. Silliman, and C. Wilson, "Pilot test of a natural-language tutoring system for physics that simulates the highly interactive nature of human tutoring," in *Artificial Intelligence in Education (16th International Conference)*, pp. 636–639, Springer, 2013.

[51] E. Goffman, "Response cries," in *Forms of Talk* (E. Goffman, ed.), pp. 78–122, Blackwell, 1981. originally in *Language* 54 (1978), pp. 787–815.

[52] R. L. Street, "Communicative styles and adaptations in physician-parent consultations," *Social Science and Medicine*, vol. 34, pp. 1155–1163, 1992.

[53] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, pp. 70–76, 2006.

[54] L. W. Barsalou, C. Breazeal, and L. B. Smith, "Cognition as coordinated non-cognition," *Cognitive Processing*, vol. 8, pp. 79–91, 2007.

[55] E. Jahr and S. Eldevik, "Response variability and turn taking in cooperative play," *Journal of Speech and Language Pathology*, vol. 2, pp. 190–194, 2007.

[56] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialog," *Behavioural and Brain Sciences*, vol. 27, pp. 169–190, 2004.

[57] H. Giles, A. Mulac, J. J. Bradac, and P. Johnson, "Speech accommodation theory: The first decade and beyond," in *Communication Yearbook 10* (M. L. McLaughlin, ed.), pp. 13–48, Sage, 1987.

[58] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[59] L. W. Barsalou, "Simulation, situated conceptualization, and prediction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, p. 1281, 2009.

[60] S. C. Levinson, "On the human 'interaction engine'," in *Roots of Human Sociality: Culture, Cognition and Human Interaction* (N. J. Enfield and S. C. Levinson, eds.), pp. 39–69, Berg, 2006.

[61] R. Dale, R. Fusaroli, N. D. Duran, and D. C. Richardson, "The self-organization of human interaction," in *The psychology of learning and motivation, Volume 59*, pp. 43–96, Academic Press, 2014.

[62] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, vol. 6, 2015.

[63] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, "Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior," in *International conference on intelligent virtual agents*, pp. 218–233, Springer, 2016.

[64] J. J. Gumperz, *Discourse Strategies*. Cambridge University Press, 1982.

[65] D. Tannen, *That's Not What I Meant! How Conversational Style Makes or Breaks Relationships*. Ballantine, 1989.

[66] N. G. Ward and Y. Al Bayyari, "American and Arab perceptions of an Arabic turn-taking cue," *Journal of Cross-Cultural Psychology*, vol. 41, pp. 270–275, 2010.

[67] P. A. Heeman, R. Lunsford, E. Selfridge, L. Black, and J. van Santen, "Autism and interactional aspects of dialogue," in *Sigdial*, 2010.