# More than words: A Framework for Describing Human-Robot Dialog Designs

James M. Berzuk
*Department of Computer Science*
*University of Manitoba*
Winnipeg, Canada
berzukj@myumanitoba.ca

James E. Young
*Department of Computer Science*
*University of Manitoba*
Winnipeg, Canada
young@cs.umanitoba.ca

***Abstract*—This paper presents a novel framework for describing human-robot interaction dialog, developed from a survey and analysis of existing systems and research. We collected data from 75 published systems and conducted an iterative thematic analysis to distill the broad range of work into key underlying factors defining them. Our framework provides a language to describe human-robot dialog systems and a new way of classifying and understanding human-robot dialog, in terms of both high-level design aspects and relevant implementation details. Our quantitative survey summary further provides a detailed, contemporary snapshot of predominant approaches in the field, highlighting opportunities for further exploration.**

***Keywords—social robotics, dialog, human-robot interaction, framework, survey***

## I. Introduction

Social robots are commonly designed to interact with people using speech as a way of simplifying communication: the robot conveys information to a person using words and a voice that a person naturally understands, and the robot listens and responds to people's utterances to leverage their existing communication ability (e.g., [1], [2]). However, despite this commonality, the field of human-robot interaction lacks a clear encompassing framework to assist designers in describing and analyzing human-robot dialog designs. As such, we conducted a scoping survey of existing published human-robot dialog systems and analyzed the data to form a framework of key design factors.

Developing frameworks to structure and explain human-robot interaction (e.g., [3]–[5]) has proven useful for supporting the analysis of various forms of interaction. For example, Kahn et al. [4] identified a series of common components within robot interaction designs, and Yanco & Drury [3] created a taxonomy for classifying overarching interaction. Frameworks for dialog systems commonly focus on aspects of the technical implementation, as in Sklar & Azhar [6], which designed an approach for robots to argue with people, or Gervits et al. [7] which developed a model for managing a robot's turn-taking behaviour. Outside of human-robot interaction, the fields of literary analysis and linguistics study communication structures to identify patterns and support analysis [8], [9]. For example, "discourse genres" (also known as "discourse modes") represent broad patterns of discourse structure, style, and function that enable one to compare a specific instance against archetypes. Our work follows this trend by developing an overarching framework specifically for human-robot dialog.

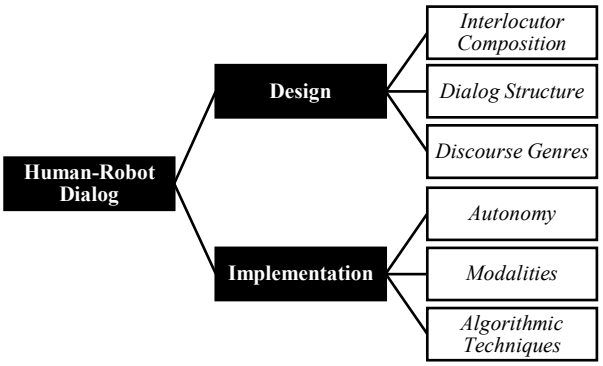


Fig. 1. Overview of our human-robot dialog framework that emerged from thematic analysis of 75 published human-robot dialog systems and designs.

We conducted a survey of human-robot interaction literature, collecting a sample of 75 unique human-robot dialog systems and designs from published works; we contacted authors as needed for additional information. Using this data, we performed an iterative, open-coding thematic analysis to extract the predominant themes and characteristics to form a framework of human-robot dialog designs (Fig. 1). This framework provides a novel encompassing method for describing and analyzing human-robot dialog systems. Further, our survey results provide a clear contemporary overview of how the community is designing human-robot dialog systems. This highlights predominant approaches and methods in human-robot dialog designs, which also points to underexplored avenues that can serve as avenues of future research for developing broader, more varied, and more natural dialog systems.

## II. Related Work

Our work fits within the tradition of developing structure and frameworks to explain human-robot interaction. For example, Hegel et al. [10] began to establish a formal definition of a social robot by examining the form and function of a system, as well as the social role and perceptions it takes on in an interaction. Another dominant early taxonomy by Yanco & Drury [3] describes human-robot interaction on dimensions including the type of task performed, the number of people and robots interacting, and the autonomy of the system's implementation. For

social robots, Bartneck & Forlizzi [11] created a descriptive framework that categorizes a design based on: robot form, communication modalities, social norms employed, autonomy, and interactivity. These works demonstrate the potential for high-level descriptive methods to direct and clarify analysis of human-robot interaction designs. We leverage such frameworks in our work, and follow this pattern to provide a similar framework specialized for human-robot dialog systems.

Kahn et al. [4] presented a finer grained approach, describing specific design patterns for social interaction such as 'The Initial Introduction,' when a robot and human exchange initial pleasantries, 'Recovering From Mistakes,' when a robot must salvage an interaction after an error, and 'Physical Intimacy,' when a robot uses physical contact to endear itself to a human. This lower-level approach is useful for considering the pieces that make up a more complete design, although it is somewhat limited in its ability to explain meta-strategies of interaction or groups of designs. Glas et al. [5] presented a framework that simultaneously identified lower-level design components, such as checking before repeating an utterance, and higher-level structural aspects, such as interaction being either 'progressive' (highly structured) or 'reactive' (immediate responses to isolated input). The authors used a data-driven approach to generate their framework, drawing from select human-robot interaction design case studies. We similarly target both high and low-level design components, and draw from existing data, developing an original human-robot dialog framework from a large selection of existing designs.

### *A. Literary Analysis*

The fields of literary scholarship and linguistics have developed numerous taxonomies [9] for classifying human communication. One approach originating in the 19th century study of rhetoric [8] is to apply archetypes to analyze broad patterns of function, style and structure within a discourse. A segment of discourse would thus be assigned to a particular 'discourse mode' [8], [12]. Traditionally, the 'discourses' this work focused on were long segments of communication, typically from a singular voice, such as passages from books or speeches [8]. We note that work occasionally uses the term 'discourse genres' instead of 'discourse modes' (e.g., [13]). Given that 'mode' already has myriad meanings in human-computer interaction, and that we adopt this technique for a novel application to human-robot dialog, for our work we use 'discourse genres' for clarity.

The field has not converged onto a universal set of discourse genres, with results instead being work-specific [9], [12]. Common traditional categories include an *argumentative* discourse, which tries to persuade the listener, a *narrative* discourse, which conveys a series of events, a *descriptive* discourse, which relates sensory information. and an *expository* discourse, which explains general information such as background details [8], [12], [14]. Identifying the discourse genre in this fashion supports comparison and analysis across bodies of text.

While discourse genres are primarily related to the intended function of a discourse, they also include stylistic and structural properties associated with that function [12]. For example, the *argumentation* genre primarily describes the goal of the discourse, to persuade someone, but also includes the style and structure that the speaker employs to communicate (such as an aggressive tone) [12]. Therefore, understanding the intent of a passage can help one understand other components such as the style used, and vice versa. It is worth noting that discourse genres are not mutually exclusive, and multiple genres can be used to describe different aspects of a discourse.

As discourse genres are typically used for analyzing oration or books [8] they cannot naively be applied to interactive, real time human-robot dialog without careful reconsideration. The standard focus on monologues makes existing discourse genres most suitable to the specialized case of a robot orating to people without reacting to or expecting any response (e.g., [15]), or a robot that merely listens and does not speak back (e.g., [16]). More commonly, robot dialog systems are designed to involve some sort of back-and-forth turn taking with shorter utterances, (e.g., [1], [17], [18]). For example, the standard and common simple query-response robot dialog designs do not fit under any traditional discourse genres; each query could be categorized differently, and the responses themselves categorized differently again. In our work we engage with the concept of discourse genres but from a human-robot interaction perspective, conducting an analysis of existing human-robot dialog designs to generate a novel set of human-robot discourse genres that can describe the functions present within an instance of robot dialog.

It is important to note that the use of discourse genres in literary and rhetorical composition has waned over time in favor of more specific, fine-grained tools. This highlights a limitation of discourse genres as being broad and thus sometimes ambiguous and less useful for targeted in-depth analyses [8]. However, discourse genres continue to be used to describe general style and structure of texts [12]. For example, recent natural language-processing algorithms have been developed to identify them [14]. Thus, discourse genres as a broad abstraction [8] are well-suited to our goal of generally describing patterns of human-robot interaction.

## III. Survey and Framework Development Method

We conducted a survey of published human-robot dialog designs and conducted an iterative open-coding thematic analysis on the results to form our novel human-robot dialog framework.

We manually and systematically searched predominant publication venues including all published issues to-date in the Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, ACM Transactions on Human-Robot Interaction, and Springer International Conference on Social Robotics. We opted to not survey commercial systems given the limited access to dialog details. We included works with any human-robot dialog design or system, even if not the main paper purpose. We liberally included any human-agent dialog communication that used a word-based medium, including on-screen text as well as speech. We excluded papers without a robot (e.g., virtual agents only) or if they were only conceptual without any implementation or prototype. Our process was to first examine the title and abstract for potential inclusion, and then exclude with deeper analysis as needed to determine eligibility. Where a system was potentially eligible, but the paper lacked sufficient detail, we contacted authors for additional details and dialog scripts (see Appendix A).

Our analysis goal was to classify works based on key interaction approaches and techniques, enabling us to both gain an overview of the current state of the field, as well as to provide a means for comparing and contrasting systems. We employed thematic analysis coding, starting with an initial set of codes drawn from literature (deductive coding, see below) and employing open inductive coding to iteratively build our code-set based on observations in the data. We kept a broad scope ranging from high-level dialog structure and function in interaction to implementation and system details that may impact dialog.

We constructed our initial code set by broadly and roughly drawing from literature: we included human-robot ratio [3], discourse genre (intended to classify the high-level structure [5] and function of the interaction), specific interaction modalities (distinguished from [11] which used a scale from monomodality to multimodality), system autonomy [3], [11], dialog interaction success, and several dimensions of implementation techniques. Starting with this set we engaged more inductive open coding, where we looked for discerning concepts relating to dialog in designs and implementations.

During analysis we continuously and iteratively updated our codes as we discovered patterns and commonalities across systems, resulting in several re-casting and re-organizations based on the fit (or lack of) of new data. Thus, the final framework ultimately emerged from both the data and the literature, as we selected dimensions and classifications to best fit the papers surveyed. Given the exploratory goal of constructing a descriptive framework (in contrast to, e.g., hypothesis testing), coding was completed by a single coder (primary author). This coder has an undergraduate degree in linguistics and is a senior undergraduate student in computer science.

## IV. Survey Results and Framework

Our process resulted in two important contributions. One, we developed a framework to describe key design features of human-robot dialog systems (Fig. 2). Two, our survey and paper classification provide the field with a contemporary overview snapshot of predominant (and rare) design approaches used (Section B, Fig. 3).

In our initial survey pass we identified 110 candidate papers. We excluded 30 that did not meet our inclusion criteria, as well 5 for redundant reporting of the same system or design. This resulted in 75 unique data points (see Appendix A).

### A. Results: Framework

Our observations converged into a framework with two broad categories (Fig. 2): *design* aspects and *implementation* aspects, and six dimensions within these.

Design aspects refer to high-level descriptors of the interaction structure and function. Our framework contains three design aspect dimensions: *interlocutor composition* – the number of robots and humans participating in an interaction, *dialog structure* – the overall flow of interaction, and *discourse genres* – the archetypical function(s) performed by the human-robot discourse.

Implementation aspects refer to concrete properties of a system's implementation that relate to dialog interaction. Our framework contains three implementation aspect dimensions: *autonomy* – the capability of the system to act without a human operator, *modalities* – the physical means by which the human and robot communicate, and *algorithmic techniques* – notable strategies for implementing the dialog system.

In the remainder of this section we detail each of these six framework dimensions.

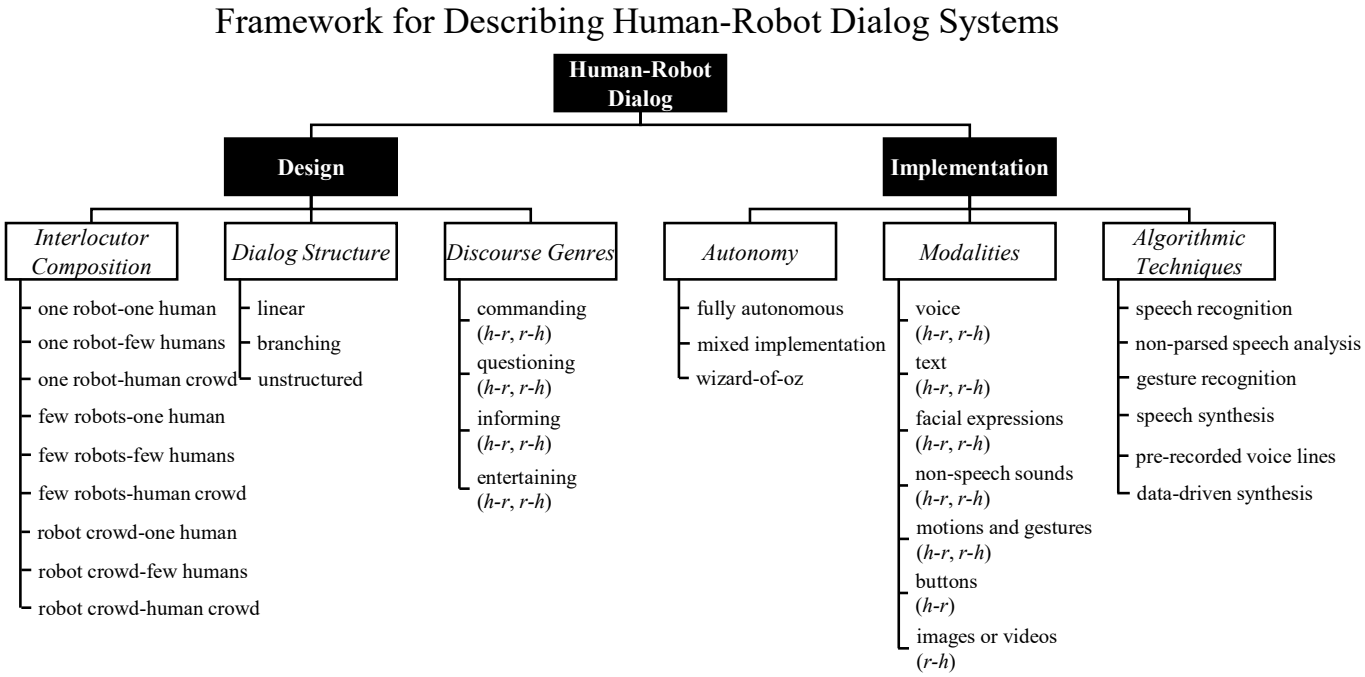


Fig. 2. The human-robot dialog framework resulting from our survey and analsysis. Note that for discourse genres and most modalities, each category includes a *human-to-robot* (*h-r*) and a *robot-to-human* (*r-h*) variant.

*1) Design Aspects: Interlocutor Composition*

*Interlocutor composition* is a measure of the number of robots and humans involved in interaction, reminiscent of the 'human-robot ratio' [3]. However, we found the exact numbers of interlocutors to be unnecessary for the purposes of understanding the form of dialog. In our analysis, we found coarse-grained group sizes to capture the important dialog-related feel and intent of systems. Thus, we categorize dialog designs as having *one*, *a few*, or *a crowd* of robots and/or humans. One key difference between *a few* and *a crowd* is that with *a few,* a robot can have one-to-one interaction with individuals, but this is much more difficult when interacting with a *crowd*, where interaction is instead directed at the crowd as whole. Designs can span compositions and change categories throughout interaction.

*2) Design Aspects: Dialog Structure*

*Dialog structure* represents the broad flow of dialog interaction over time and its relative linearity. At one extreme dialog can be fully pre-scripted and *linear*, with perhaps small deviations or loops. Alternatively, dialog can be somewhat more reactive and *branching* based on input or responses (if primarily pre-scripted). At the other extreme a dialog can be completely reactive and *unstructured*, based heavily on immediate user input; these typically consist of largely isolated, reactive utterances. This dimension is loosely inspired by Glas et al.'s 'progressive' and 'reactive' interaction flows [5], and draws a distinction between designs that are highly predictable versus those that have more interaction uncertainty and broader possibilities. It is possible for a design to fall into multiple categories, or change categories, as interaction evolves.

*3) Design Aspects: Discourse Genres*

*Discourse genre*s, inspired by linguistics and literary analysis [8], [12], are qualitative categories that capture functional intent of dialog, focusing on what a dialog design was trying to accomplish. Our analysis resulted in four functions of discourse. *Commanding* refers to encouraging another party(s) to do something, whether by direct instruction or a more polite suggestion. *Questioning* is the bilateral, turn-taking process of asking questions and receiving answers. *Informing* is when information is supplied unilaterally, in a less interactive or directing manner. *Entertaining* includes a variety of actions with the primary goal of entertaining, such as acting, singing, or playing a game.

Each of these discourse functions can be performed *by* a person (or robot), or *to* a person (or robot), where the meaning of the interaction differs greatly in each case; thus each function results in two distinct discourse genres. For example, a robot commanding a person has quite a different meaning (regarding system and interaction design) compared to a person commanding a robot. Discourse may further differ in delivery structure and style, in addition to the basic functional distinction.

Note that a dialog design may incorporate several discourse genres, for different components or as dialog passes through phases. For example, a design may simultaneously employ both *human-questioning-robot* and *robot-commanding-human*, as the robot responds to a human's requests for information, but as it does so also attempts to tell the human to take some action.

*4) Implementation Aspects: Autonomy*

We noted the importance of a dialog implementation being *fully autonomous*, fully *Wizard-of-Oz'd*, or some combination of autonomy and remote operation (e.g., for complex parts of behavior), *mixed implementation*. This implementation information is helpful for understanding the potential naturalness of the interaction, as well as what resources may be required to reproduce it. While this is similar to categorizations found in prior frameworks ([3], [11]), we found that additional granularity was not helpful for understanding dialog.

*5) Implementation Aspects: Modalities*

A system's interaction modalities, tied to the implementation technologies used, is highly relevant for understanding how and why dialog was designed. Most modalities were used by both people and robots, such as *voice*, *text* (displayed by the robot or typed by the human), *facial expressions*, *non-speech sounds* (e.g., robot beeps or human clapping), and *motions or gestures*. However, there were human- or robot-specific modalities: we noted the use of *buttons* for human-to-robot interaction, and displaying *images or videos* for robot-to-human interaction.

*6) Implementation Aspects: Algorithmic Techniques*

Our analysis revealed several predominant *algorithmic techniques* that we included based on high-level relevance for dialog design; these dictated what was possible or feasible. For input processing we found *speech recognition* (speech to text), *non-parsed speech analysis* (e.g., volume or tonal analysis), and *gesture recognition*. For output generation we found *speech synthesis* (text-to-speech), *pre-recorded voice lines*, and *data-driven synthesis* (machine learning).

### *B. Results: Frequency Data*

As detailed in Appendix A, our survey and analysis resulted in the classification of all works found. We present a summary overview of these results, organized by the categories of the framework in Fig. 3. This provides a contemporary snapshot of what techniques are commonly (as well as less commonly) explored in the community by date.

### *C. Discussion*

We analyzed a corpus of 75 published human-robot dialog designs and systems, and distilled them into a broad framework that explains the main components of dialog design, as well as relevant implementation details that influence the dialog design. The resulting framework is able to cleanly categorize and explain human-robot dialog in six simple, primary dimensions that can be used to facilitate analysis of existing systems, comparison between systems, as well as consideration for new designs.

The summary frequency statistics resulting from our survey (Fig. 3) provide a contemporary snapshot for understanding how dialog systems are designed and built in the community today.

Our results indicate that many dialog designs are focused on the robot performing actions, and not the human. For all four basic discourse functions, the robot-to-human genre was more frequent in surveyed systems than the corresponding human-to-robot genre. While this makes sense, for example, with a robot performing, we note that it is striking that it is more common for

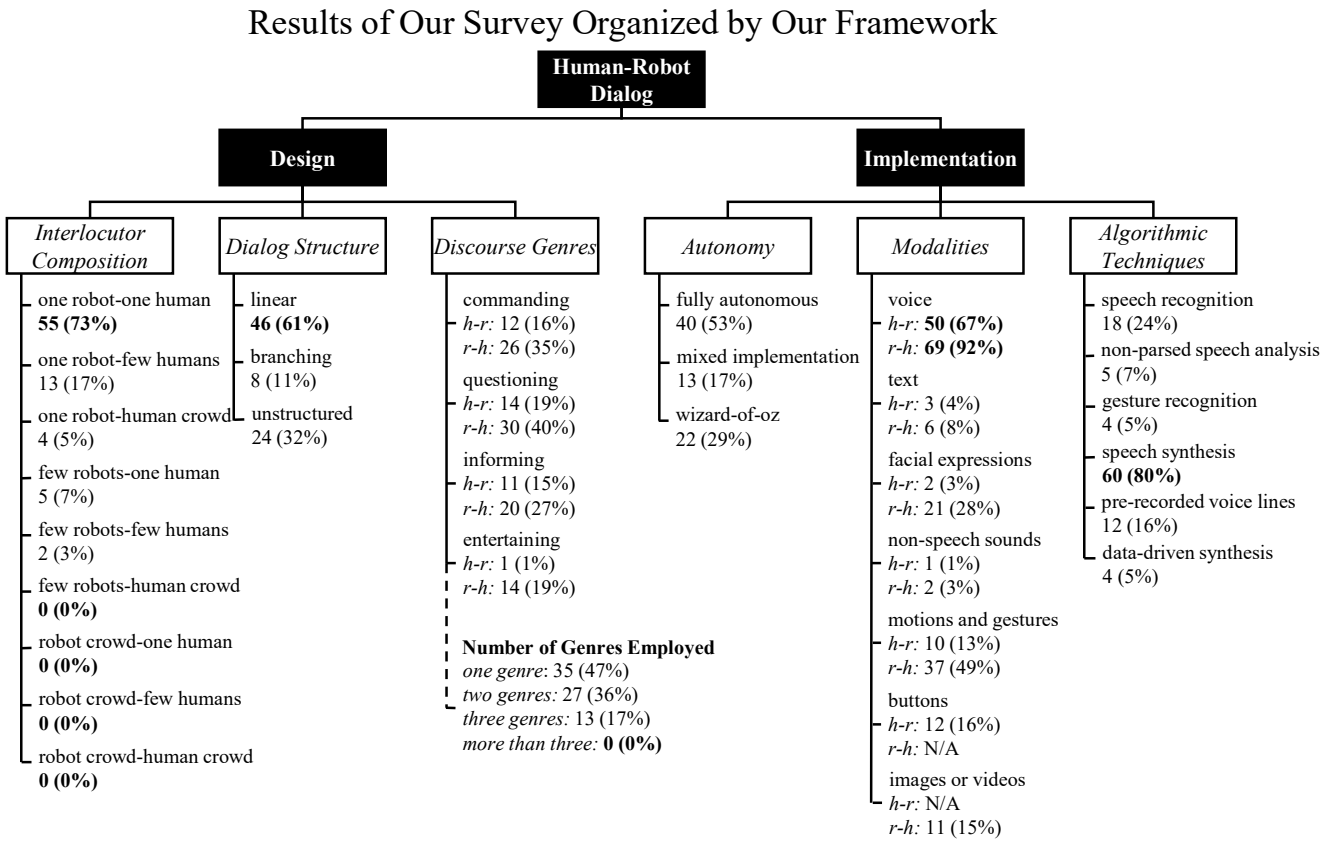


Fig. 3. The results of our survey organized by the framework we developed, where numbers represent how many surveyed systems or designs (of the 75) that satisfied that category. Note that in each dimension other than autonomy, a system can be included in more than one category (percentages may not add to 100%).

research to investigate a robot telling a person what to do, than reacting to or listening to a person. Perhaps this discrepancy can be considered more deeply by the field.

Along a similar vein, we note that 72% of all systems follow a highly linear or pre-scripted branching interaction. On the surface, this makes sense from an implementation perspective: such systems are simply easier to build than more complex unstructured interaction flow. However, this points to a potential deeper problem. From a user-centered design perspective, we would expect a designer to plan a dialog structure based on specific interaction goals. For example, perhaps the dialog system should be flexible and forgiving (e.g., in a daycare setting), or rigid and exact (e.g., a kiosk). However, our data suggests that it may be implementation simplicity, and not necessarily interaction needs, driving many research systems. We note that this is reflected in common dialog toolkits (e.g., such as with the SoftBank NAO and Pepper NaoQi systems) which–by design–steer designers toward keyword recognition-and-response systems.

As is to be expected for dialog, the bulk of research involves voice-based communication, both from the human and from the robot. However, although people use copious amounts of facial expressions and other gestures while talking, we only found this in a minority of robotic designs. Further, of particular interest is novel forms of dialog-related interaction made possible by technology, including the use of images, videos, and buttons. This highlights the new dialog possibilities available in human-robot interaction that do not exist in dialog between people.

Over half (53%) of systems were fully autonomous, with only 29% completely reliant on a Wizard operator. We found this encouraging, as it demonstrates the substantial effort being given toward practical autonomous systems. On a related note, there was a strong tendency for systems to use speech synthesis, perhaps to support flexible interaction and generated, responsive text, even though recorded voice would sound more natural. Conversely, speech recognition was only used in a about a quarter of cases, possibly given its error-prone nature. We suspect that these observations are correlated, in that the high number of autonomous systems is enabled by design decisions that avoid problematic solutions such as speech recognition.

We found clear opportunities for new research directions in our interlocutor composition results. Most systems, by far, are dyads of one robot and one person, with nearly all systems using one robot with any composition of humans. Perhaps this stems from the expense of robots (e.g., to develop a crowd), but working with groups of people and robots remains a relatively unexplored area in terms of dialog design.

There was one outlier observation under discourse genres, where we only found one instance of people entertaining robots. Häring et al. [19] conducted a study where two robots played a game with a human, in which all players entertained each other. Perhaps future work could explore a robot acting as an audience

member for a person practicing a performance art (e.g., as with the gig simulator in the movie "Tenacious D in the Pick of Destiny" [20]). The social presence and dialog structure could provide positive and negative feedback to shape the practice performance. Similarly, *human-informing-robot* systems such as Park et al. [16]'s robot that a child tells a story to, could be modified for the robot to have a requirement or stated goal of being entertained.

In all, we found our resulting framework to be a useful mechanism for examining the current state of the field relating to human-robot dialog. Further, the numerical results from our survey and classification highlighted important trends in design and opportunities for further exploration.

## V. Case Studies

We conducted a set of case studies to illustrate the potential use of our framework as a vocabulary to explain and contrast human-robot dialog systems. We selected five candidate systems from those surveyed systems as recent instances that span the framework categories. We present each system using the framework and discuss and contrast them along the dimensions.

Morimoto et al. [1] developed a customer service robot to receive customers' complaints. The humanoid robot (a Robovie2) would process a customer by listening to their complaints, asking questions to clarify the situation, and then apologizing and offering an explanation to address their concerns. The details of the system are described in Fig. 4. It has a *linear* dialog structure as while the interaction involves some limited branching and loops as necessary, it flows through a linear set of phases, always ending up in the same place. The interaction spans three separate discourse genres: *human-informing-robot*, where the robot listens to the customer's complaints; *robot-informing-human*, where the robot addresses the concerns; and in the paper's proposed model, *robot-questioning-human*, where the robot asks the customer for more information and context. Examining the interaction through the lens of these discourse genres highlights its bidirectional, conversational nature. While the robot ultimately guides the interaction according to its set phases, both parties have lengthy periods within the interaction in which they are the leading speaker.

Vilk & Fitter [15] developed a comedian robot (a SoftBank NAO) to perform in front of crowds, described in Fig. 5; it uses the *robot-entertaining-human* discourse genre. While the robot did employ *non-parsed speech analysis* to adjust the endings of its jokes based on the volume of the crowd's response, its dialog structure is classified as *linear* as this does not result in any deeper branching. The framework classification makes it clear at a glance that this interaction is nearly entirely driven by the robot, with the human crowd serving simply as an audience.

Mizumaru et al. [21] developed a robot, described in Fig. 6, that aims to mimic a security guard and correct the behaviour of pedestrians on the street. The robot (a Robovie-R3 on a mobile base) would approach pedestrians who were using a smartphone

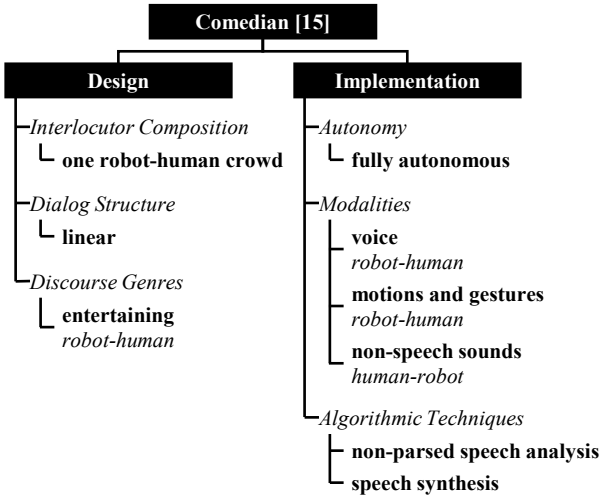


Fig. 5 Vilk & Fitter [15]'s comedian robot that performed a *one robot* stand-up routine in front of a *human crowd*. It *entertained* the audience with its *linear* routine, and was *fully autonomyous* throughout. The robot delivered its performance through *synthesized voice* as well as *motions*, and reacted to the volume of the *sounds* from the crowd, which it analyzed using *non-parsed speech analysis*.

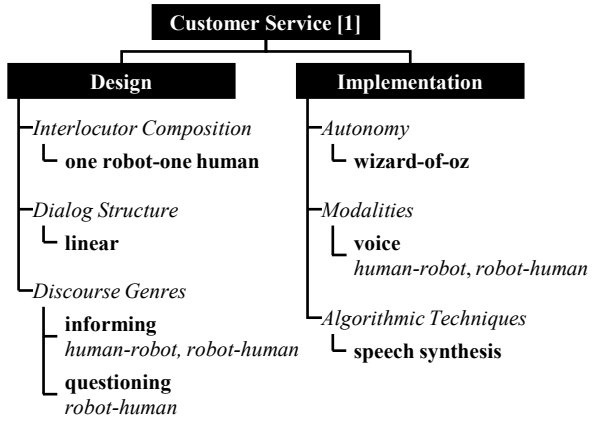


Fig. 4 Morimoto et al. [1]'s customer service robot that listened to and addressed customer complaints. This robot design used the *one-robot-one-human* interlocutor composition and a *linear* dialog structure to *inform* and *question* the customer, and to be *informed* by the customer in return, all controlled by a *wizard-of-oz* operator. They communicated using *voice*, with the robot generaing its using *speech synthesis*.

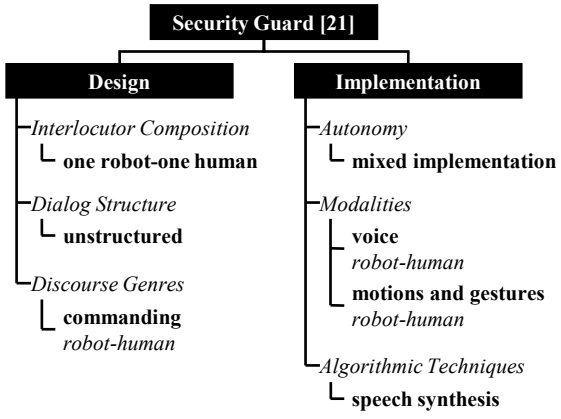


Fig. 6 Mizumaru et al. [21]'s security guard robot that discouraged passerbys from using their phones while walking. With an interlocutor composition of *one robot-one human*, it *commanded* them to stop with a single, *unstructured* utterance. With some of its decisionmaking left to a human operator, it used synthesized voice as well as motion to persuade its targets.

while walking and instruct them to stop, citing the potential danger. Its dialog structure is described as *unstructured*, rather than *linear* or *branching*, as it only made a single, short utterance to each pedestrian. We can see that this interaction is completely unilateral as it has no human-robot modalities. The agency of the human participant then is limited simply to whether to comply with the robot's instruction.

Park et al. [16] developed a pair of robots, described in Fig. 7, that acted as listeners for children to tell stories to. The robots (two Tegas) would sit in front of a single young child and listen to them as they told a story, backchanneling throughout. While one of the robots did not process the child's speech at all, the other used *non-parsed speech analysis* to tailor its backchanneling to the situation. Viewing the interaction through the lens of the framework quickly shows it is quite one-sided, and in this case led by the human side. Examining the list of modalities underscores this, but also reveals that the robot listeners are still active participants in the interaction despite not speaking themselves.

Alves-Oliveira et al. [22] developed a robot, described in Fig. 8, to lead students through group learning scenarios. The robot (a SoftBank NAO) was designed to guide a group of students as they played a multiplayer educational video game. Its dialog diverged considerably depending on the students' actions in the game, meaning the interaction had a *branching* structure. While the discourse genre shifted throughout the interaction between *commanding*, *questioning*, and *informing*, all three were directed from the robot to the human. This emphasizes the way in which the robot teacher was driving the interaction, with the students generally acting in response.

Each of the selected systems demonstrates a very different type of human-robot dialog interaction, and this is highlighted by the framework. One stark difference we have showcased is between the interaction that is more evenly balanced between the human and robot parties ([1]) and the other four which are more heavily controlled by one party or the other, to varying degrees. This is not the only area of distinction that this lens can

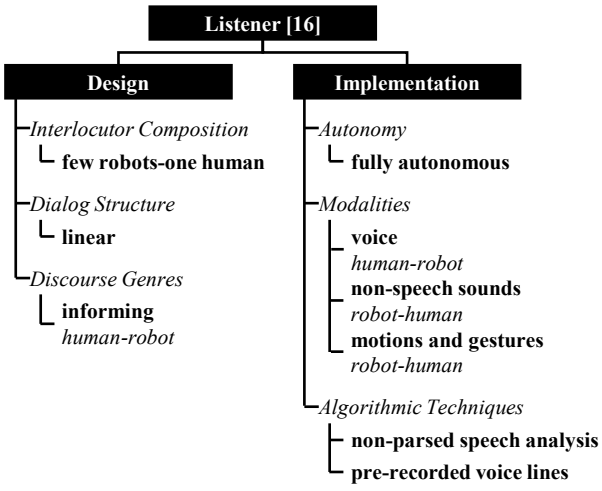


Fig. 7 Park et al. [16]'s pair of robots deisgned to a child tell them stories. The *few robots* were *informed* by *one human* child who delivered a *linear* story using *voice*. The *fully autonomous* robots backchanneled using *pre-recorded sounds* and *motions*, and one of them tailored this to the child using *non-parsed speech analysis*.

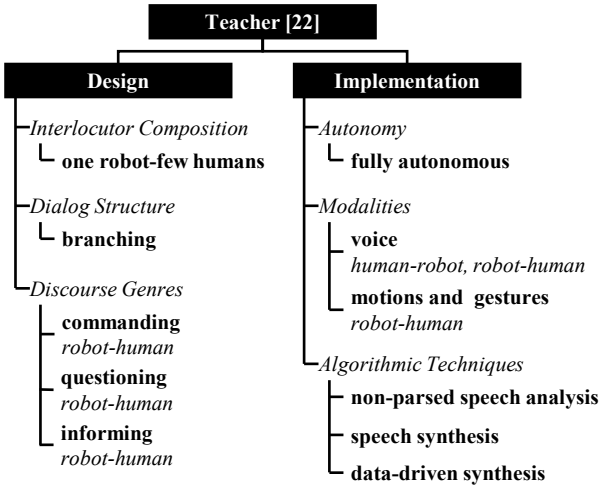


Fig. 8 Alves-Oliveira et al. [22]'s robot designed to teach students how to play a video game. The *one robot* taught a *few human* students by *commanding* them with instructions, *informing* them about the game, and *questioning* them to test their understanding. The robot was *fully autonomous*, and its dialog *branched* according to the students' behaviour. The robot communicated using *synthesized voice* lines as well as *motions*, and the students communicated back using *voice* which the robot evaluated using *non-parsed speech analysis*.

highlight. For example, three of the interactions ([15], [16], [21]) were very focused on a single purpose, and thus each fit into a singular discourse genre, while the other two ([1], [22]) were more varied and thus spanned three genres each. Another clear distinction is between the unstructured, one-off instructions of [21] compared to the larger, more structured interactions in the others. The framework we have developed provides a systematic way to identify and discuss these differences in interaction designs.

## VI. Limitations and Future Work

A primary limitation of our work is our focus on published systems in the academy. There are several commercial systems, and proprietary software (e.g., Pepper installations) which may not follow the patterns we see here. In addition, expanding our survey to include additional academic venues would bring in more systems that may enable more detailed analysis.

An important future direction for this framework is ongoing analysis of how the various dimensions relate to one another. For example, we did note that 75% of *one robot-human crowd* systems employed the *robot-entertaining-human* discourse genre, in comparison to only 19% of systems overall. However, we found such a cross-dimension synthesis to be largely infeasible in our case due to the limited number of systems available and surveyed, in contrast to the many possible cross-dimension combinations. This makes it difficult to generalize about the interactions between different dimensions of the framework: as case in point, our above observation relies on only four total *one robot-human crowd* systems surveyed. To support initial analysis and developing directions for future inquiry we have included a cross-tabulation of our survey data in Appendix B.

Performing these sorts of comparisons is further complicated by the vast array of contexts that dialog systems are employed in. The relationships between the dimensions when applied to

systems in, for example, a healthcare context may differ greatly from those in an educational setting. While high-level observations may be made, more informed conclusions will still require a closer analysis of the systems involved.

We had initially included in our survey the success of an interaction design to perhaps identify which combinations of factors are more likely to produce success. Unfortunately, we were not able to find a systematic way to evaluate the success of the systems. The papers we surveyed used disparate measures of success, with some not measuring dialog success at all or conducting any relevant evaluation. Such analysis would be more feasible with a smaller, more focused group of systems where their effectiveness can be more deeply examined.

We had also considered focusing more closely on the technical implementations of the systems. Many of these points would manifest as additional categories under the dimension of algorithmic techniques, such as whether the system simply acted in response to predefined keywords versus performing more detailed lexical analysis, or whether the interaction rigidly followed a predefined script. For this initial survey we focused more on the higher-level design aspects, leaving these lower-level algorithmic details for future work.

Finally, we note that sometimes the boundary between categories can be ambiguous. For example, if a robot is supplying information in a playful manner, it could be difficult to draw a clean line between *informing* and *entertaining* when assigning discourse genres. However, as described in Section 2a an overarching framework uses necessarily broad abstractions to describe a wide variety of disparate texts. As has happened in linguistics, finer granularity would likely require more targeted frameworks.

## VII. Conclusions

We developed an in-depth framework for human-robot dialog, developed from an analysis of a corpus of 75 existing designs and systems in the research literature. This framework provides a novel and broad vocabulary and set of dimensions for describing and analyzing human-robot dialog. Further, our survey results provide an important contemporary snapshot of prominent work in the field, which highlights trends in research as well as future directions for exploration.

Overall, we envision that this work will provide a backbone for ongoing human-robot dialog research, and hope that our framework serves as a tool for researchers to better understand their work and explore new directions.

## Acknowledgment

This project was funded by the Natural Sciences and Engineering Research Council of Canada, through their Discovery Grants program. This project was further supported by the University of Manitoba Faculty of Science Undergraduate Student Research Award.

We would like to thank all of the authors who cooperated with us in conducting this survey and who went out of their way to dig up difficult-to-find data and study details. Without all their support this work would not have been possible.

## References

[1] D. Morimoto, J. Even, and T. Kanda, "Can a robot handle customers with unreasonable complaints?," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 579–587, doi: 10.1145/3319502.3374830.

[2] D. Leyzberg, A. Ramachandran, and B. Scassellati, "The Effect of Personalization in Longer-Term Robot Tutoring," *ACM Trans. Human-Robot Interact.*, vol. 7, no. 3, Dec. 2018, doi: 10.1145/3283453.

[3] H. A. Yanco and J. Drury, "Classifying human-robot interaction: an updated taxonomy," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 2004, vol. 3, pp. 2841–2846, doi: 10.1109/ICSMC.2004.1400763.

[4] P. H. Kahn *et al.*, "Design patterns for sociality in human-robot interaction," in *HRI 2008 - Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction: Living with Robots*, 2008, pp. 97–104, doi: 10.1145/1349822.1349836.

[5] D. F. Glas, T. Kanda, and H. Ishiguro, "Human-robot interaction design using interaction composer: Eight years of lessons learned," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2016, vol. 2016-April, pp. 303–310, doi: 10.1109/HRI.2016.7451766.

[6] E. I. Sklar and M. Q. Azhar, "Argumentation-Based Dialogue Games for Shared Control in Human-Robot Systems," *J. Human-Robot Interact.*, vol. 4, no. 3, p. 120, Dec. 2015, doi: 10.5898/jhri.4.3.sklar.

[7] F. Gervits, R. Thielstrom, A. Roque, and M. Scheutz, "It's About Time: Turn-Entry Timing For Situated Human-Robot Dialogue," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 86–96, Accessed: Sep. 29, 2021. [Online]. Available: https://aclanthology.org/2020.sigdial-1.12.

[8] R. J. Connors, "The Rise and Fall of the Modes of Discourse," *Coll. Compos. Commun.*, vol. 32, no. 4, p. 444, 1981, doi: 10.2307/356607.

[9] M. Fludernik, "Genres, Text Types, or Discourse Modes? Narrative Modalities and Generic Categorization," *Style*, vol. 34, no. 2, pp. 274–292, 2000.

[10] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, "Understanding Social Robots," in *2009 Second International Conferences on Advances in Computer-Human Interactions*, Feb. 2009, pp. 169–174, doi: 10.1109/ACHI.2009.51.

[11] C. Bartneck and J. Forlizzi, "A design-centred framework for social human-robot interaction," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 2004, pp. 591–594, doi: 10.1109/roman.2004.1374827.

[12] C. S. Smith, *Modes of Discourse*. Cambridge University Press, 2003.

[13] W. F. Hanks, "Discourse Genres in a Theory of Practice," *Am. Ethnol.*, vol. 14, no. 4, pp. 668–692, Oct. 1987, Accessed: Sep. 08, 2021. [Online]. Available: http://www.jstor.org.uml.idm.oclc.org/stable/645320.

[14] W. Song, D. Wang, R. Fu, L. Liu, T. Liu, and G. Hu, "Discourse mode identification in essays," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Jul. 2017, vol. 1, pp. 112–122, doi: 10.18653/v1/P17-1011.

[15] J. Vilk and N. T. Fitter, "Comedians in cafes getting data: Evaluating timing and adaptivity in real-world robot comedy performance," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 223–231, doi: 10.1145/3319502.3374780.

[16] H. W. Park, M. Gelsomini, J. J. Lee, and C. Breazeal, "Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2017, vol. Part F1271, pp. 100–108, doi: 10.1145/2909824.3020245.

[17] A. Weiss *et al.*, "Robots asking for directions — The willingness of passers-by to support robots," in *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, 2010, pp. 23–30, doi: 10.1109/hri.2010.5453273.

[18] J. Peltason, N. Riether, B. Wrede, and I. Lütkebohle, "Talking with robots about objects: A system-level evaluation in HRI," in *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2012, pp. 479–486, doi: 10.1145/2157689.2157841.

[19] M. Häring, D. Kuchenbrandt, and E. André, "Would you like to play with me? How robots' group membership and task features influence human-robot interaction," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2014, pp. 9–16, doi: 10.1145/2559636.2559673.

[20] L. Lynch, *Tenacious D in The Pick of Destiny*. United States: New Line Cinema, 2006.

[21] K. Mizumaru, S. Satake, T. Kanda, and T. Ono, "Stop Doing it! Approaching Strategy for a Robot to Admonish Pedestrians," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2019, vol. 2019-March, pp. 449–457, doi: 10.1109/HRI.2019.8673017.

[22] P. Alves-Oliveira, P. Sequeira, F. S. Melo, G. Castellano, and A. Paiva, "Empathic Robot for Group Learning: A Field Study," *ACM Trans. Human-Robot Interact.*, vol. 8, no. 1, Mar. 2019, doi: 10.1145/3300188.