ACL Lifetime Achievement Award

Smart Enough to Talk With Us? Foundations and Challenges for Dialogue Capable AI Systems

Barbara J. Grosz
Harvard University
Cambridge, MA 02138

1. Beginnings and Background

I am deeply grateful for the honor of this award, all the more so for its being completely unexpected. I am especially pleased by the recognition this award gives to our early attempts to build computational models of dialogue and to develop algorithms that would enable computer systems to converse sensibly with people. ACL was my first academic home. I presented my first paper, the paper that laid out the basics of a computational model of discourse structure, at the 1975 ACL meeting. Then, after several decades of research centered on dialogue systems, my research focus shifted to modeling collaboration. This shift was driven in part by the need for computational models of collaborative activities to support dialogue processing, a topic which I explore briefly below, and in part by limitations in speech processing and semantics capabilities. Research in these areas has advanced significantly in the last decade, enabling advances in dialogue as well, and I have recently returned to investigating computer system dialogue capabilities and challenges. I am glad to be back in my intellectual home.

The use of language has been considered an essential aspect of human intelligence for centuries, and the ability for a computer system to carry on a dialogue with a person has been a compelling goal of artificial intelligence (AI) research from its inception. Turing set conversational abilities as the hallmark of a thinking machine in defining his "imitation game," more commonly referred to as "The Turing Test." In the 1950 *Mind* paper in which he defines this game, Turing conjectures thus, "I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted." (Turing 1950, p. 442). Though the Turing Test remains an elusive (and now debatable) goal, this conjecture has proved true. Natural language processing research has made great progress in the last decades, with personal assistant systems and chatbots becoming pervasive, joining search and machine translation systems in common use by people around the world. Although it is commonplace to hear people talk about these systems in anthropomorphic terms — using "she" more frequently than "it" in referring to them — the limitations of these systems' conversational abilities frequently leads people to wonder what they were thinking or if they even were thinking. As the dialogue in Figure 1 illustrates, at root these systems lack fundamental dialogue capabilities. The first assignment in a course I have been teaching, "Intelligent Systems:

doi:10.1162/COLI_a_00313

B: Where is the nearest gas station?
S: [list of 16 nearby gas stations]
B: Which ones are open?
S: Would you like me to search the web for "Which ones are open?"

Figure 1
A typical problematic dialogue with a personal assistant.

Design and Ethical Challenges," is to test phone-based personal assistants. Although some systems might handle this particular example, they all fail similarly, and the range of dialogue incapabilities students have uncovered is stunning. As I have argued elsewhere, the errors that systems make reveal how far we still have to go to clear Turing's hurdle or to have systems smart enough to (really) talk with us (Grosz 2012).

The first generation of speech systems, which were developed in the 1970s, divided system components by linguistic category (e.g., acoustic-phonetics, syntax, lexical and compositional semantics), and dialogue challenges of that era included identifying ways to handle prosody and referring expressions, along with nascent efforts to treat computationally language as action. In contrast, current spoken language research takes a more functional perspective, considering such issues as sentiment analysis, entrainment, and deception detection. Typical current dialogue challenges include such functions as turn-taking, clarification questions, and negotiation. The database query applications of earlier decades, which lacked grounding in dialogue purpose, have been replaced by template filling and chat-oriented applications that are designed for relatively narrowly defined tasks.

Various findings and lessons learned in early attempts to build dialogue systems have potential to inform and significantly improve the capabilities of systems that currently aim to converse with people. Although the computational linguistic methods that have recently enabled great progress in many areas of speech and natural-language processing differ significantly from those used even a short while ago, the dialogue principles uncovered in earlier research remain relevant. The next two sections of this paper give brief summaries of the principal findings and foundations in models of dialogue and collaboration established in the 1970s through the 1990s. The following section examines principles for dialogue systems derived from these models that could be used with current speech and language processing methods to improve the conversational abilities of personal assistants, customer service chatbots, and other dialogue systems. The paper concludes with scientific and ethical challenges raised by the development and deployment of dialogue-capable computer systems.

2. Foundations: From Sentences to Dialogue

My first papers on dialogue structure (Grosz [Deutsch] 1974, 1975; Grosz 1977) were based on studies of task-oriented dialogues I collected to inform the design of the discourse component of a speech understanding system. In these dialogues, an expert (who was eventually to be replaced by the system) instructed an apprentice in equipment assembly. The expert and apprentice were in separate rooms, they communicated through a teletype interface (with a camera available for still shots on demand), and the apprentice was led to believe the expert was a computer system. These dialogues were collected as part of what were, to my knowledge, the first "Wizard of Oz" experiments.[1]

1 The moniker "Wizard of Oz" for such experiments was coined only several years later by a group at Johns Hopkins University.

(1) E: First you have to remove the flywheel.
(2) A: How do I remove the flywheel
(3) E: First loosen **the two small Allen head setscrews holding it to the shaft**, then pull **it** off.
(4) A: OK
Subdialogue about the two setscrews.
(5) A: **The two setscrews** are loose, but I'm having trouble getting the wheel off.
(6) E: Use ***the wheelpuller***. Do you know how to use it?
(7) A: No …
(8) E: Loosen **the screw in the center** and place the jaws around the hub of the wheel, then tighten **the screw**.

Figure 2
Evidence of structure in "naturally occurring" dialogues.

Figure 2 is one example from this collection. It illustrates the influence of dialogue structure on definite noun phrase interpretation. In Utterance (3), the expert (E) directs the apprentice (A) to loosen two setscrews. In Utterance (8), the expert instructs the apprentice to loosen "the screw in the center". The only intermediate explicit mention of screw-like objects are subsequent references to the setscrews in Utterance (5) and the elided subdialogue about them between Utterances (3) and (5). With only two objects, one cannot be in the center. Indeed, the screw to which the expert refers in Utterance (8) is in the center of the wheelpuller. This dialogue fragment has a structure that parallels the task of removing the flywheel. When the first step of that task (loosening the setscrews) is completed, as indicated by the apprentice in Utterance (5), the focus of attention of expert and apprentice move to the step of pulling off the wheel with the wheelpuller. At the point of mention of "the screw in the center", the setscrews are no longer relevant or in context; the wheelpuller is.

A more striking example is provided by the dialogue sample in Figure 3. Each of the successful assemblies of the air compressor ended in one of the ways shown in this figure. Despite admonitions of grammar teachers and editors that pronouns be used to refer to the last object mentioned that matches in number and gender, actual use of pronouns can vary markedly. In this case, the pronoun "it" is used to refer to the air compressor after half an hour of dialogue in which there was no explicit mention of it.

Again, context and the focus of attention has shifted over the course of the assembly and the expert-apprentice dialogue related to it. At the point at which the last utterance occurs, the air compressor is assembled and is the focus of attention. Not only did the participants in these dialogues have no trouble understanding what was to be plugged in or turned on (even though they were in a room full of equipment that might have provided other options), but also readers of the transcripts of these dialogues are not confused.

E: Assemble **the air compressor**.
30 min. later, with no intervening explicit mention of the air compressor.
E: Plug **it** in. / See if **it** works.

Figure 3
The initial, most striking examples of dialogue structure influencing referring expressions.

John came by and left the **groceries**.
Stop that you kids.
And I put **them** away after he left

Figure 4
Intonation and Discourse Structure.

The narrative fragment in Figure 4, which is from the work of Polanyi and Scha (1983), shows that this kind of discourse structure influence on referring expressions is not restricted to task dialogues. The linearly nearest referent for the pronoun "them" in the last utterance is the children mentioned in the previous utterance. It is the groceries that were put away though, not the children. This example also illustrates the role of prosody in signaling discourse structure. Intonation is crucial to getting the interpretation of this story right. The second utterance is said as an aside, with strong intonational markers of this separation. Prosody, then, is integral to dialogue processing, and spoken dialogue cannot be handled as a pipeline from speech recognition to pragmatic plan recognition.

These fragments illustrate a primary finding of my early dialogue research: dialogues are not linear sequences of utterances nor simply question–answer pairs. This work established that task-oriented dialogues are structured, with multiple utterances grouping into a dialogue segment, and their structure mirrors the structure of the task. Subsequently, Candy Sidner and I generalized from task dialogues to discourse more generally, defining a computational model of discourse structure (Grosz and Sidner 1986). This model defines discourse structure in terms of three components as shown in the schematic in Figure 5. The linguistic structure comprises the utterances themselves, including their various linguistic features, which cluster into dialogue segments. Just as written language is broken into paragraphs, the utterances of spoken language naturally form groups. The attentional state tracks the changes in focus of attention of the dialogue participants as their conversation evolves. The intentional structure comprises the purposes underlying each dialogue segment and their relationships to one another. Relationships between dialogue segment purposes (DSPs) determine embedding relationships between the corresponding discourse segments (DS), and thus they are foundational to determining changes in attentional state. These three components of discourse structure are interdependent. The form of an utterance and its constituent phrases are affected by attentional state and intentional structure, and they in turn may engender changes in these components of discourse structure.

Sidner and I examined two levels of attentional state. The global level, which is portrayed in Figure 5, is modeled by a focus space stack. Focus spaces contain

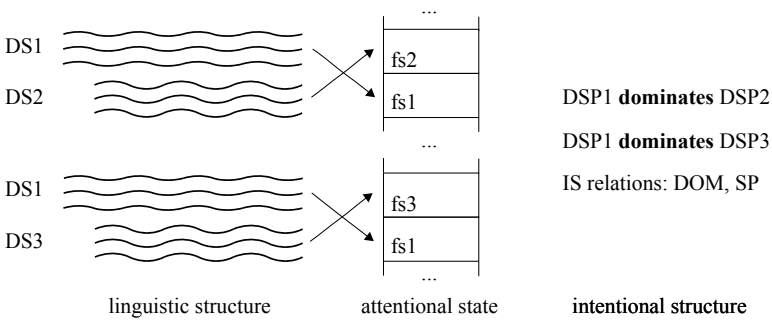


Figure 5
The tripartite model of discourse structure.

representations of the objects, properties, and relations salient within a discourse segment as well as the discourse segment purpose. When a new dialogue segment is started, a space is pushed onto the stack, and when the segment completes, the corresponding focus space is popped from the stack. This level of attentional state influences definite noun phrase interpretation and generation and intention recognition processes. Access to objects, properties, and relations at lower levels of the stack may not be available at higher levels. The local level of attentional state, which Sidner and I referred to as "immediate focus" ("local focus" being a tongue twister), models smaller shifts of attention within a discourse segment. Sidner (1979, 1981, 1983a) deployed immediate focusing as a means of controlling the inferences required for interpreting anaphoric expressions like pronouns and demonstratives by limiting the number of entities considered as possible referents. Webber's contemporaneous work on the range of anaphors a single seemingly simple definite noun phrase could yield made clear the importance of being able to limit candidates (Webber 1979, 1986). It is important to note that "focus" has been used for various other ideas in linguistics. For instance, the Prague School of computational linguistics uses topic/focus information structures to relate sentences to discourse coherence (Hajicova 2006), and "topic" is used to refer to speech signal properties in the prosody literature and to syntactic properties in some grammatical theories.

A short while later, Joshi and Weinstein investigated ways to control a system's overall reasoning by using information about which entity was central in an utterance. They defined a concept of "centers of a sentence in a discourse," which bore strong similarity to immediate focus. They used this concept to derive initial results on how differences in centers changed the complexity of inferences required to integrate a representation of the meaning of an individual utterance into a representation of the meaning of the discourse of which it was a part (Joshi and Weinstein 1981). A year after their paper appeared, I was fortunate to be invited by Joshi to be a visiting faculty member at the University of Pennsylvania. During that visit Joshi, Weinstein, and I merged their centering ideas with Sidner's discourse ideas into what became known as "centering theory," which highlighted the process of "centering of attention" and considered constraints on generation as well as interpretation (Grosz, Joshi, and Weinstein 1983; Grosz, Weinstein, and Joshi 1995). We adopted centering terminology because of the already widely varying uses of "focus." A vast literature on centering has followed, including cross-linguistic empirical and psycholinguistic investigations of centering predictions in English, German, Hebrew, Italian, Japanese, Korean, Turkish, and many other languages; variations in ways centers are realized in language; and development of a range of centering algorithms. Recently, Kehler and Rohde have developed a probabilistic model that combines centering with coherence-relation driven theories (Kehler and Rohde 2013).

Investigations of linguistic markers of discourse boundaries soon followed. Hirschberg and Pierrehumbert launched research in this area with their influential work on the intonational structuring of discourse. They showed that pitch range, amplitude, and timing correlated with global and local structures defined in the tripartite model Sidner and I had developed (Hirschberg and Pierrehumbert 1986; Pierrehumbert and Hirschberg 1990). Hirschberg and Litman (1987, 1993) showed that pitch accent and prosodic phrasing distinguish between discourse and sentential usage of such cue phrases as "now." Their example from a radio talk show, "So in other words I will have to pay the full amount of the uh of the tax now what about Pennsylvania state tax?" illustrates the importance of determining which use a speaker intended. Later Hirschberg and I investigated empirically ways in which speakers use intonation to mark discourse

structure boundaries (Grosz and Hirschberg 1992). For her Ph.D. research, Nakatani worked with us on determining ways in which accent placement varied with grammatical function, form of referring expression, and attentional status (Hirschberg, Nakatani, and Grosz 1995; Hirschberg and Nakatani 1996).

One challenge for these investigations of intonational markers of discourse structures was to obtain examples from multiple people of spontaneous speech with essentially the same content. In one of the earlier attempts to develop a corpus of spontaneous, naturally occurring dialogue, we designed a set of tasks which led multiple speakers to give similar directions around the Cambridge/Boston area. To connect intonation with discourse structure, the resulting Boston Directions Corpus was annotated by multiple trained annotators for prosodic events using the ToBI labeling conventions (Hirschberg and Beckman; Beckman, Hirschberg, and Shattuck-Hufnagel 2004) as well as for discourse structure using a guide based on the tripartite model of discourse structure. The design of materials that can evoke spontaneous speech of a near identical nature from multiple speaker remains a challenge, one that urgently needs to be addressed for the improvement of capabilities of personal assistant and chatbot systems.

3. Intentional Structure and Formalizing Collaborative Plans for Dialogue

The third component of discourse structure, the intentional structure, has roots in work on language as action in philosophy (Austin 1962; Grice 1969; Searle 1969) and subsequent work on language as planned behavior in artificial intelligence (Bruce 1975; Cohen and Perrault 1979; Allen and Perrault 1980; Sidner 1983b). The example in Figure 6, adapted from Bruce (1975), provides an example of the need to infer the intentions behind an utterance to respond appropriately. In this dialogue fragment, C is attempting to get B to get rid of the bugs on the rhubarb plant, not completely successfully. I have used this example because it is less task-oriented than others and so illustrates the general need for this capability. This research on language as planned behavior yielded several algorithms for recognizing speakers' intentions and generating utterances that communicated them. This work, though, was largely focused on individual utterances or pairs of utterances, and the planning methods it used were based on AI research on single agent planning.

Sidner and I attempted to generalize these methods to the dialogue setting, but eventually determined that intentional structure could not be represented by any sum of single agent plans. This realization led us to formulate a computational formalization of collaborative two-agent plans, which we called SharedPlans (Grosz and Sidner 1990). Kraus and I subsequently generalized this initial model to one able to handle an arbitrary number of agents and more complex recipes for actions (Grosz and Kraus 1996, 1999). Hunsberger in his Ph.D. dissertation subsequently modified the initial formalizations to enable proving properties of SharedPlans (Hunsberger 1998) and to formally integrate decision making (Grosz and Hunsberger 2006). Lochbaum's

C: The rhubarb has holes.
B: So?
C: It's covered with little bugs
B: We should use vegetable friendly spray to get rid of them

Figure 6
A small example of language as planned behavior (adapted from B. Bruce, 1975).

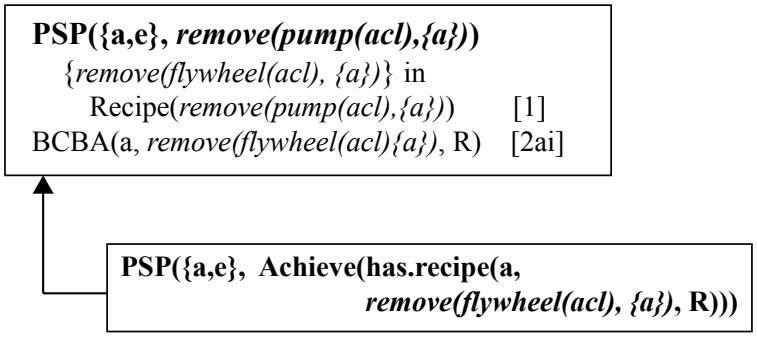
dissertation research showed how SharedPlans could be used as the basis of the intentional structure of dialogue (Lochbaum 1998). In recent years, supporting evidence for the importance of collaboration to language processing has come from research in brain sciences and developmental psychology that has established the importance of social interaction to brain development and language learning.

Figure 7 provides English glosses of the major elements of the SharedPlans specification of collaboration. The term "recipe" in (2) refers to (knowing) a way to do an action. In the formal SharedPlans specification, each of the requirements for commitment is represented in terms of intentions and those for consensus in terms of mutual belief. There can be no intention without ability: Philosophical and logical theories of intention require that an agent be able to do any action it forms an intention to do. As a result, the commitments stipulated in various parts of the specification implicitly carry such a requirement. The literature on teamwork demonstrates that each element is required for teamwork to succeed by showing ways in which teamwork could fail if the requirements on agent beliefs or intentions that it specifies are not met. For instance, if some team members do not commit as in (1), those team members could walk away from the activity without meeting their obligations or neglect to inform other team members if they were unable to meet their obligations; if they do not commit to each other's success as in (5), they could fail to provide help to each other. In such cases, the teamwork could fail. The interactions among the intentions and (mutual) beliefs in this specification are an embodiment of the need for the capabilities for collaboration to be designed into systems from the start (Grosz 1996). Alternative computational theories of teamwork (Cohen and Levesque 1991; Sonenberg et al. 1992, inter alia) differ in the particulars, but have similar types of principles, which if not adhered to cause the teamwork to fail.

Each of the requirements or constraints in the SharedPlan specifications might engender extended dialogue related to satisfying it. The dialogue in Figure 2 can be used to sketch the way Lochbaum (1998) used SharedPlans to model intentional structure and to illustrate the kind of reasoning about intentions that is needed to recognize dialogue structure and understand utterances in a dialogue. In essence, the intentional structure is a collection of SharedPlans (some partial, some completed) and relationships between these plans (e.g., that one is needed as part of another) determine relationships between the DSPs constituting the structure. As the schematic in Figure 8 shows, the expert and apprentice have a (partial) SharedPlan (PSP) to remove the pump, which includes as a substep removing the flywheel (i.e., doing so is part of the recipe). Clause [2ai] in the figure is part of the logical specification for element (4) in the SharedPlan specification

1. Each team member **commits** to the team's performance of the group activity.
2. Team members must reach **consensus on a (high-level) "recipe"**; in SharedPlans, recipes may be partial, revised over time.
3. Team members must reach **consensus on allocation of (subtasks)**, taking into account agents' capabilities.
4. Team members **commit** to assigned subtasks.
5. Team members **commit** to each others' success.

Figure 7
SharedPlans model of collaboration.



Utterances (3)-(4) are understood and produced in this context

Figure 8
Analysis of the first subdialogue.

in Figure 7; it stipulates that the apprentice must believe s/he is able to remove the pump. This stipulation leads the expert and apprentice to have a subsidiary (partial) SharedPlan for the apprentice to know how to perform the flywheel removal action (i.e., have the recipe for doing so). Utterances (3) and (4) in the dialogue relate to this subsidiary SharedPlan. This clarification subdialogue corresponds to one type of knowledge precondition for action, the requirement of knowing the recipe for an action to be able to do that action. Another type of clarification dialogue can result from the knowledge precondition that an agent needs to be able to identify the objects that participate in an action; the elided subdialogue about setscrews in Figure 2 is just such a knowledge-precondition clarification subdialogue. With Utterance (5), the apprentice closes off these subdialogues and opens a new one related to pulling off the flywheel. The focus spaces containing representations of the setscrews are popped from the stack so that only the flywheel and, after Utterance (6), the wheelpuller are in focus when Utterance (8) occurs. They are the objects relevant to determining the referent of "the screw in the center."

4. From Theories to Design Principles

Even in this era of big data, machine learning, and efforts to build end-to-end dialogue systems, theoretical work on dialogue can provide guidance for systems design and testing. Interestingly, in a different AI domain, Shoham has argued persuasively for the usefulness of logically sound theories for design of interactive systems (Shoham 2015). A major challenge in applying the theories of dialogue structure and collaboration I have described in current dialogue systems, though, is that such systems typically will not have the kinds of detailed planning and recipe knowledge that Lochbaum's approach required. Allen's work on TRAINS (Allen et al. 1995) and Traum's negotiation dialogue systems (Traum et al. 2008) have developed ways to limit the amount of domain knowledge needed. I will digress briefly to show an alternative, design-oriented approach to deploying the principles these theories establish in dialogue systems. To do so, I need to sketch the way we have used the theory of collaboration in a similarly plan-knowledge poor setting to yield design principles for information sharing in the context of our work on health care coordination (Amir et al. 2013).

There are multiple ways one can deploy the kind of theoretical frameworks these theories provide. They can serve to generate system specifications or portions of them,

as represented by Kamar's Ph.D. research on interruption management (Kamar, Gal, and Grosz 2009; Kamar 2010; Kamar, Gal, and Grosz 2013). They can guide design, as they have in recent research on collaborative interfaces for exploratory learning environments (Gal et al. 2012; Amir and Gal 2013; Uzan et al. 2015; Segal et al. 2017). They can also provide an analytic framework for understanding behavior before design or for performance analysis. Our health care work deployed SharedPlans in these last two ways. In work with Stanford pediatricians, we are developing methods to improve information sharing among the 12–15 care providers typically involved in meeting the ongoing heath care needs of children with complex conditions.

Formative interviews with parents of such children, physicians, and therapists revealed several characteristics of teamwork in this setting that not only raise challenges for coordination, but also make for a poor fit with prior AI planning methods (Amir et al. 2015). Based on these characteristics, we dubbed this kind of teamwork FLECS, because the team has a **f**lat structure (no one is in charge); the team's activities are **l**oosely coupled, have **e**xtended duration, and are **c**ontinually being revised; and team members typically act in a **s**yncopated manner, not synchronously. As a result, team members have very little of the detailed knowledge of each other's plans required by value-of-information approaches to determining what information to share. Figure 9 shows a mapping between SharedPlans constraints, designated by italics, and the kinds of capabilities they suggest a system should provide to a health care team in light of the extreme locality of information about tasks in FLECS teamwork. This analysis led Amir in her Ph.D. research to develop an interaction-based algorithm to support information sharing for the FLECS teamwork settings (Amir, Grosz, and Gajos 2016).

For systems to reply appropriately to people and in a way that makes sense in an extended dialogue, it will be crucial to get the structure and intentions right. Doing so will matter for linguistic form and for content. Figure 10 enumerates some possible ways to use SharedPlans principles to ameliorate flaws in the behavior of current dialogue systems that are based on machine learning of various sorts. The top list shows some ways in which dialogue structure can derive from participants' decision making and from their activities. The second list shows ways in which participants' commitments, abilities, and needs might give rise to dialogue structure. Regardless of a dialogue system's underlying implementation methods, incorporating into the design of these methods some knowledge of dialogue structure and collaboration could improve the system, just as adding knowledge of syntactic structure or semantic relations seems to help squashing algorithms perform better (Peng, Thomson, and Smith 2017), and

SP challenge for information sharing without information overload: extreme locality of information about delegated tasks

SP: Consensus on recipe: Provide support for providers establishing agreement on high-level approach and establishing mutual belief.

SP: Recipes may be partial and evolve over time: Support dynamically evolving plans.

SP: Team members commit to performance of group activity and to each other's success: Support communication and coordination at appropriate levels and times.

Figure 9
Analytic use of SharedPlans theory for health care coordination.

Decision making and participant activities (physical or mental):
Purposes of the dialogue interaction: What are dialogue participants doing? Why are they talking?
Consensus on recipe: How are they carrying out the activities directed toward their purposes?
Consensus on subtasks: Who is doing what?

Commitments that lead to structures spawned by participants' abilities and needs:
to group activity
to assigned subtasks
to each other's success

Figure 10
Potential uses of SharedPlans theory for dialogue systems.

hierarchical reinforcement learning (Georgila, Nelson, and Traum 2014) overcomes some of the incoherence of dialogue systems based on flat reinforcement learning. Encouraging results of recent work on incorporating centering-like information into generative models for coreference resolution in text processing (Ji et al. 2017) also provides support for pursuing this research direction.

5. Scientific and Ethical Challenges for Dialogue Systems

Two themes have guided my dialogue research from its start: people matter and language itself matters. The characteristics of dialogue that I have described raise challenges for current systems. They are the reason that open domains (e.g., social chatbots, Twitter) are harder to handle than closed domains (travel assistants, customer service) and that short dialogues (Q/A pair) are easier than extended dialogue. Real human dialogues, though, are typically extended, and people frequently stray outside of any closed domain as their needs shift from those expected by system designers. To build systems capable of conversing sensibly with people requires close observation and analysis of ways people talk with one another, collecting data in real situations, and celebrating the full glory of language use rather than building systems that require people to control their language use. The extraordinary progress in computational linguistics recently argues for being more ambitious in our scientific and engineering goals. Some challenges I hope researchers will take on include:

How can we integrate the fundamental principles of dialogue and collaboration models with data-driven machine-learning approaches?

How can we build generalizable models that transfer from one domain to another, given the handcrafting required by semantic-grammar based systems and the large amounts of dialogue data machine learning methods would require (e.g., for incorporating dialogue purpose in their models)?

How can we meet the societal needs of handling dialogue across the full range of natural languages and different cultures?

How can researchers help ensure that research results are better integrated into deployed systems, so that they do not make avoidable errors, potentially causing ethical problems?

Computational linguistics and natural-language processing systems also raise some of the most serious ethical challenges at the current moment. In addition to the potential harm of dialogue system failure, these challenges include the effect of social chatbots on ways people communicate with one another and of systems inadequately performing jobs like customer service costing people in time and effort. A principle that is derivable from the intentions included in the SharedPlans specification is that agents cannot commit to doing an activity that they know they are incapable of doing. This principle is one every dialogue system should obey, but few current systems do. The dialogue fragment in Figure 1 is annoying or humorous, depending on one's perspective. Other kinds of mistakes made by current dialogue systems, though, are not just awkward or annoying, but raise ethical concerns. Students in the course I mentioned earlier have found serious, potentially life-threatening errors in their investigations of personal assistant systems and in toys that claim to be dialogue capable. For instance, based on experiences of talking with each other, people naturally generalize from individual instances of language capability. So, we would expect that a system able to answer the question "Where is the nearest ER?" would also be able to provide answers to "Where can I get a flu shot?" and "Where can I go to get a sprained ankle treated?". Unfortunately, there are systems that fail on such seemingly related questions. Providing a list of web pages that describe how to treat a sprained ankle, as one system students tested did, may not be so serious for a sprained ankle, but a similar failure could be disastrous were the question about a heart attack or other serious medical condition.

Just as capabilities for collaboration cannot be patched on but must be designed in from the start (Grosz 1996), so too ethics must be taken into account from the start of system design. So, I will end by asking readers to think about the data they use, the systems they build, and the claims they make about those systems from an ethical perspective, and to do so from the initial stages of their formulating the design of these systems.

6. Concluding Thanks and A Bit of Personal History

I thank ACL, the people who nominated and recommended me for this honor, and the selection committee. I am awed to join the company of earlier awardees, a list that includes many people whose work I have greatly admired from my earliest days in computational linguistics. Eva Hajicova, Aravind Joshi, and Karen Sparck Jones, in particular, taught me a great deal about language structure and function, and their work and advice helped shape my research on dialogue.

Alan Kay inspired my move from theoretical areas of computer science to AI, by suggesting as a thesis topic the challenge of building a system that would read a children's story and then re-tell it from one character's point of view. (This challenge remains open so far as I know.) Around the time I realized that children's stories, full as they are of cultural and social lessons, were more difficult to understand than ordinary discourse, Bill Paxton and Ann Robinson lured me to the SRI speech understanding group. The challenge they posed to me was to build the discourse and pragmatics components of the SRI speech system. Everyone involved in the early 1970s speech systems efforts recognized that their systems needed to take context into account, but no one had yet formulated ways to operationalize context in a natural-language system. Bill and Ann argued (correctly) that task-oriented dialogue was simpler than children's stories and (incorrectly, at least at the time) that speech was easier than text because it carried more information. They along with Jane Robinson and Gary Hendrix gave me a natural-language processing platform on which I could explore various approaches

to dialogue models, and they taught me a great deal about AI and linguistics, neither of which I had formally studied. Led by Don Walker, the SRI speech group was an amazingly supportive research environment, one that inspired my approach to leading my own research group, and also to building computer science efforts at Harvard.

My research has been enriched greatly by collaborations with Candy Sidner, a partner in developing the tripartite model of discourse structure and the initial SharedPlans model of collaborative plans; with Julia Hirschberg, whose work convinced me the time had become right for integrating dialogue models with speech and intonation; and with Sarit Kraus, a partner in generalizing SharedPlans and investigations of teamwork. The students, postdoctoral fellows, and younger colleagues with whom I have worked have enabled the research advances described in this paper, inspired me, and enriched my academic life. I have been fortunate throughout these years to work with superb Harvard students, undergraduate and graduate, and also with students based in other universities whose interest in dialogue or in collaboration led them to ask me to join their dissertation committees and even to become unofficial advisers. I thank them all for their science and for the joy they brought to our investigations.

In the talk on which this paper is based, I also revealed some lessons I learned as my career progressed, which I have tried to pass on to my own students. I include brief mention of them here in hopes they will encourage subsequent generations of graduate students and other early career stage researchers to tackle hard, important problems, and to persist in investigating novel approaches unpopular though they may seem at first. One lesson is not to be discouraged too easily by the skepticism of senior researchers, who may have their own biases and be wrong. When he heard about my thesis research, Noam Chomsky remarked that it was an interesting problem but advised I would never succeed because dialogue could not be formalized. John McCarthy told me an understanding of people's cognitive processes was irrelevant to AI, and I failed to convince him that it did matter if one was interested in language use. A second lesson is to persist despite negative feedback from peers about the importance of a problem. When Sidner and I first presented SharedPlans, arguing that AI models of planning for individual agents would not suffice for modeling collaborations, several AI planning and reasoning researchers suggested we just think harder. Some of them later developed their own teamwork models, having realized we were right. Einstein said "To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science." (Einstein and Infeld 1938). With the increasing prevalence of computer systems communicating in natural languages and being used to analyze human language for all manner of features, the importance of addressing the deepest problems of language and communication has become even greater than it was when I first ventured down the path of discovering ways to operationalize context.

References

Allen, James F. and C. Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178.

Allen, James F., Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, and David R. Traum. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

Amir, Ofra and Yakov Kobi Gal. 2013. Plan recognition and visualization in exploratory learning environments. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(3):16.

Amir, Ofra, Barbara Grosz, and Krzysztof Z. Gajos. 2016. Mutual influence potential networks: Enabling information sharing in loosely-coupled extended-duration

teamwork. In *Proceedings of the International Joint Conference for Artificial Intelligence (IJCAI'16)*.

Amir, Ofra, Barbara J. Grosz, Krzysztof Z. Gajos, Sonja M. Swenson, and Lee M. Sanders. 2015. From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*, pages 1419–1428.

Amir, Ofra, Barbara J. Grosz, Edith Law, and Roni Stern. 2013. Collaborative health care plan support. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, pages 793–796.

Austin, John Langshaw. 1962. *How to do things with words*. Oxford University Press.

Beckman, Mary E., Julia Hirschberg, and Stephanie Shattuck-Hufnagel. 2004. The original ToBI system and the evolution of the ToBI framework. In S. A. Jun, editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, Oxford University Press, pages 9–54.

Bruce, Bertram C. 1975. Generation as a social action. In *Proceedings of the 1975 workshop on Theoretical Issues in Natural Language Processing*, pages 64–67.

Cohen, Philip R. and Hector J. Levesque. 1991. Teamwork. *Noûs*, 25(2):487–512.

Cohen, Philip R. and C. Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212.

Einstein, Albert and Leopold Infeld. 1938. *The Evolution of Physics*. Cambridge University Press.

Gal, Ya'akov, Swapna Reddy, Stuart M. Shieber, Andee Rubin, and Barbara J. Grosz. 2012. Plan recognition in exploratory domains. *Artificial Intelligence*, 176(1):2270–2290.

Georgila, Kallirroi, Claire Nelson, and David R. Traum. 2014. Single-agent vs. multi-agent techniques for concurrent reinforcement learning of negotiation dialogue policies. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 500–510.

Grice, H. Paul. 1969. Utterer's meaning and intention. *The Philosophical Review*, 78(2):147–177.

Grosz, Barbara and Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. In *Second International Conference on Spoken Language Processing*, pages 429–432, Banff, Canada.

Grosz, Barbara J. 1977. The representation and use of focus in a system for understanding dialogs. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 67–76, Cambridge, MA.

Grosz, Barbara J. 1996. Collaborative systems (AAAI-94 presidential address). *AI Magazine*, 17(2):67–85.

Grosz, Barbara J. 2012. What question would Turing pose today? *AI magazine*, 33(4):73–81.

Grosz, Barbara J. and Luke Hunsberger. 2006. The dynamics of intention in collaborative activity. *Cognitive Systems Research*, 7(2):259–272.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, pages 44–50.

Grosz, Barbara J. and Sarit Kraus. 1996. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357.

Grosz, Barbara J. and Sarit Kraus. 1999. The evolution of sharedplans. In A. Rao and M. Woolridge, editors, *Foundations of Rational Agencies*, Kluwer Academic Press, pages 227–262.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, Barbara J. and Candace L. Sidner. 1990. Plans for discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communications*, MIT Press, Cambridge, MA, pages 417–444.

Grosz, Barbara J., Scott Weinstein, and Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz [Deutsch], Barbara. 1974. The structure of task-oriented dialogs. In *IEEE Symposium on Speech Recognition: Contributed Papers, Pittsburgh: Carnegie Mellon University*, pages 250–254.

Grosz [Deutsch], Barbara. 1975. Establishing context in task-oriented dialogs. *American Journal of Computational Linguistics*.

Hajicova, Eva. 2006. ACL lifetime achievement award: Old linguists never die, they only get obligatorily deleted. *Computational Linguistics*, 32(4):457–469.

Hirschberg, Julia and Mary Beckman. The ToBI annotation conventions. Available at `http://www.cs.columbia.edu/~julia/files/conv.pdf`.

Hirschberg, Julia and Diane Litman. 1987. Now let's talk about now. In *Proceedings of the Association for Computational Linguistics*, pages 163–171, Stanford, CA.

Hirschberg, Julia and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hirschberg, Julia and Christine Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the Association for Computational Linguistics*, pages 286–293, Santa Cruz, CA.

Hirschberg, Julia, Christine H. Nakatani, and Barbara J. Grosz. 1995. Conveying discourse structure through intonation variation. In *ESCA Workshop on Spoken Dialogue Systems*, pages 189–192, Visgo, Denmark.

Hirschberg, Julia and Janet Pierrehumbert. 1986. The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 136–144.

Hunsberger, Luke. 1998. Making SharedPlans more concise and easier to reason about. In *Proceedings. International Conference on Multi Agent Systems*, pages 433–434.

Ji, Yangfeng, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. Dynamic entity representations in neural language models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1839, Association for Computational Linguistics, Copenhagen, Denmark.

Joshi, Aravind K. and Scott Weinstein. 1981. Control of inference: Role of some aspects of discourse structure-centering. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI'81)*, pages 385–387.

Kamar, Ece. 2010. Reasoning Effectively Under Uncertainty for Human-Computer Teamwork. Ph.D. thesis, Harvard University.

Kamar, Ece, Ya'akov Gal, and Barbara J. Grosz. 2009. Incorporating helpful behavior into collaborative planning. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 875–882.

Kamar, Ece, Ya'akov Kobi Gal, and Barbara J. Grosz. 2013. Modeling information exchange opportunities for effective human–computer teamwork. *Artificial Intelligence*, 195:528–550.

Kehler, Andrew and Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, 39(1-2):1–37.

Lochbaum, Karen E. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.

Peng, Hao, Sam Thomson, and Noah A. Smith. 2017. Deep multitask learning for semantic dependency parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2048, Association for Computational Linguistics, Vancouver, Canada.

Pierrehumbert, Janet and Julia Bell Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*, MIT press, pages 271–311.

Polanyi, Livia and Remko J. H. Scha. 1983. The syntax of discourse. *Text-Interdisciplinary Journal for the Study of Discourse*, 3(3):261–270.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*, volume 626. Cambridge University Press.

Segal, Avi, Shaked Hindi, Naomi Prusak, Osama Swidan, Adva Livni, Alik Palatnic, Baruch Schwarz, and Ya'akov (Kobi) Gal. 2017. Keeping the teacher in the loop: Technologies for monitoring group learning in real-time. In *International Conference on Artificial Intelligence in Education*, pages 64–76. Springer.

Shoham, Yoav. 2015. Why knowledge representation matters. *Commun. ACM.*, 59(1):47–49. 10.1145/2803170.

Sidner, Candace Lee. 1979. Towards a computational theory of definite anaphora comprehension in English discourse, Massachusetts Institute of Technology, Cambridge, Artificial Intelligence Lab.

Sidner, Candace L. 1981. Focusing for interpretation of pronouns. *Computational Linguistics*, 7(4):217–231.

Sidner, Candace L. 1983a. Focusing and discourse. *Discourse Processes*, 6(2):107–130.

Sidner, Candace L. 1983b. What the speaker means: The recognition of speakers' plans in discourse. *Computers & Mathematics with Applications*, 9(1):71–82.

Sonenberg, E., G. Tidhar, E. Werner, D. Kinny, M. Ljungberg, and A. Rao. 1992. Planned team activity. *Artificial Social Systems*, 890.

Traum, David, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents*, pages 117–130, Springer, Berlin.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Uzan, Oriel, Reuth Dekel, Or Seri, and Ya'akov (Kobi) Gal. 2015. Plan recognition for exploratory learning environments using interleaved temporal search. *AI Magazine*, 36(2):10–21.

Webber, B. 1986. So what can we talk about now? In Barbara J. Grosz, Karen Sparck-Jones, and Bonnie Lynn Webber, editors, *Readings in Natural Language Processing*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 395–414.

Webber, Bonnie Lynn. 1979. *A Formal Approach to Discourse Anaphora*. Routledge.