Empirical Methods in Discourse: Limits and Prospects

Richmond H. Thomason thomason@isp.pitt.edu
Pamela W. Jordan jordan@isp.pitt.edu
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260

Abstract

We claim that the methods used in approaches to discourse all display in rather acute form the gap between theory and evidence that is common in many areas of cognitive science. We propose enhancing the existing approaches with a simulation-based approach combining computational work in generation and interpretation, and discuss some of the issues in the design of a simulation architecture.

Problem statement

Wherever there is data there are empirical methods. But to yield understanding—including the sort of understanding needed to guide action—these methods need to be properly coupled with a well informed body of theory. In cognitive science, it has not always been easy to find ways to combine the two productively.

Discourse provides a good example of a domain in which the current methods that have evolved in a number of disciplines (cognitive psychology, communications, sociology, linguistics, computer science, philosophy of language) leave something to be desired, despite the interesting and important work that has been carried out in all of these fields.

Survey of empirical methods in discourse

Many different fields come to grips with discourse; all of them use methods that are broadly empirical. But this leaves room for a great deal of variation in methods.

Cognitive psychology. The best work in this area combines interesting, high-level hypotheses about discourse with psychological experiments using human subjects. The experiments are replicable, and the results are quantifiable. But the relation between the experimental results and the hypotheses is loose; it is hard to see how the experimental methods could be used to test issues about processing methods or discourse structure that are controversial in AI.

Sociology, sociolinguistics. There is a robust tradition of dealing with real data, there are rigorous methods of transcription and data collection. But the theoretical dimension is in general shallow. The hypotheses that can be tested by the methods occupy a very low level on the cognitive scale.

Philosophy of language. The philosophical tradition is not as active as it once was, but it has served as an inspiration for much of the later work. The empirical side of this tradition is weak; philosophers never really followed J.L. Austin's advice of using the OED as a research tool. Examples are constructed and the work tends to concentrate on a small number of simple, artificial cases. Even the best work fails to produce satisfactory detailed explanations of even simple phenomena.

Linguistics. The methods that linguists feel comfortable with do not work well in discourse; world knowledge and domain reasoning are not matters of language. In areas that linguists have addressed (such as presupposition), accommodation is a problem—the core rules can be flouted by speakers. When nonhyphenated linguists have worked on discourse, they have tended to use methods inherited from syntax; examples are constructed, and the theories try to explain patterns of well-formedness or associations of meanings with syntactic structures.

Computer science. Probably the most exciting work going on in discourse these days is being carried out by computer scientists. The computational work imposes additional constraints having to do with the need to build systems. Partly because of this, it is in the computational tradition that one finds the most systematic attempts to relate world knowledge and domain reasoning to discourse. But the theoretical and empirical methods used by the computational community are mainly inherited from other disciplines, and within computer science you see divisions between theory-related work and corpus-related work that are similar to the ones dividing disciplines like philosophy and

sociolinguistics. Fundamental theoretical issues (e.g. "Are there such things as discourse relations?") are unresolved in the computational arena, and it is hard to see how to use actual evidence to resolve them.

Diversity

There is neither a strong, dominant theoretical paradigm nor a favored empirical methodology that is common to the above disciplines. The theoretical ideas and empirical methods vary widely, and are often borrowed from other areas: game theory, logic, ordinary language philosophy, cognitive psychology, sociology, artificial intelligence. To some extent, this borrowing is quite legitimate: cognition, general purpose reasoning and language come together in discourse, and to do justice to all these aspects one has to use appropriate methods for dealing with them. But borrowing the methods is not enough, since one thing that makes these separate fields work is the boundaries they have drawn to focus on specialized problems. Providing a unified theory that deals with the appropriate problems of discourse is more than borrowing a bit here and a bit there; you somehow have to do it without the boundaries.

Comparison with Alan Newell's "twenty questions" problem

It is useful to compare the situation in discourse with the one in cognitive psychology on which Allen Newell commented in (Newell 1973). He begins by saying that he feels half enthusiastic and half puzzled and frustrated by the state of research at that time. Newell's enthusiasm had to do with a growing body of experimental results. There was an empirical methodology that was shared by a large, talented community, involving experimental design and statistical analysis of quantitative data. There was general agreement on computational theoretical models of cognition. The empirical methods produced a large body of generalizations that have held up pretty well; see the survey in (Newell 1990, Chapter 1). Newell's frustration in 1973 was due to the looseness of contact between the experimental results and the relevant theory. To escape the dilemma, he proposed the development of a broader, more tightly knit body of simulatable theory.

A Case Study of the Problem

As soon as the analysis of discourse gets beyond the shallowest area, the need grows to ascribe hidden intentions to the participants. Consider, for instance, the following text, discussed in (Fox 1987).

1. Bee: 'hmhhh 'hh So you gonna be around this weekend?
2. Ava: Uh::m. (0.3) Possibly
3. Bee: Uh it's a four day weekend-I have so much work to do it isn't funn//y.
4. Ava: Well, tomorrow I have to go in.

Fox analyzes (1) as interpreted as a pre-invitation by Ava in (2), which is later, at utterance (3), reinterpreted as a request for information. The hesitation in (2) is cited as evidence for the initial interpretation.

However, there is no conclusive way to support this interpretation. As Fox points out, although hesitation typically signals a pre-rejection, the hesitation could signal that Ava is thinking about her plans. Even if it were possible to interrupt a dialog like this and ask the subject to describe her inferences at this particular point, the subject may herself not be able to articulate the inferences she just made; these inferences are often not the focus of a speaker's attention, and even when they are made an interruption might well mask them. If the subject is asked the question later, the inferences may well be forgotten.

We have been exploring a design for task-oriented dialogs that asks the subjects to build a graphical representation of the design as it unfolds. Though, of course, this graphical representation does not reflect everything the subject is thinking, the design has proved highly useful in studying aspects of the collaborative process; see (Jordan & Moser 1995).

However, we believe that more is needed in order to relate evidence about discourse to higher-level representations and reasoning strategies.

Suggestion: model conversation by simulating agents that can interpret and generate

The crucial part of our own proposal is somewhat like Newell's: try to simulate complete conversational agents. (The agents may be simple-minded, but they have to incorporate domain reasoning, generation, and interpretation.) These simulations will put us in a better position to test theoretical hypotheses about conversation; we will provide some examples of this later.

The design also permits a more or less objective test of "uptake." Each agent will maintain a view of what has been established in the course of the conversation. In successful, well coordinated conversations, the participating agents will leave the conversation with the same view of what has been achieved.

There are many differences between this idea and Newell's. (1) We do not insist on a discourse architecture that is suitable for general cognitive simulation. (2) We are not necessarily interested in simulating the fine structure of human performance. (We assume that an agent that simulates a "rational reconstruction" of conversational skills will perform well enough as a first approximation.) (3) There is no well developed body of quantitative generalizations to which we need to try to fit the performance of the system. (4) There is no generally accepted methodology for collecting and refining

empirical data about conversation; so part of the goal of the simulation is to help to create this methodology.

There are a number of ways to motivate our proposal from current lines of research. Here is one, based on recent theoretical work by Asher, Oberlander, and Lascarides.[1]

Their technique is broadly linguistic: they use intuitions about naturalness in constructed dialogs to test hypotheses about default discourse inferences. E.g., in the discourse "John pushed Max; Max fell" the relevant intuition is that two events occurred, in the order in which they are invoked in the discourse. Hypotheses that explain this inference are modified (usually, by invoking overriding defaults) by testing them against other constructed examples.

But this leaves the hypotheses rather underdetermined, since when a hypothesis is refuted, there is no reliable way to tell whether to give it up and replace the originally posited default with another, or to retain the default and invoke an exception to it. The problem is that the hypotheses are defeasible, rather than indefeasible as in syntax and semantics, where a counterexample leaves no doubt about whether generalizations have to be retracted.

A simulation environment would provide a better way of testing these hypotheses. You can look at simulated behavior as a function of (among other things) the defaults of the agents, and the environment would enable you to test the effects on discourse of changing the defaults, while holding other factors constant. In case incoherent discourse is generated or coherent discourse is incorrectly interpreted, the simulation program may let you track the fault to a specific default, and it may be possible to predict to some extent what changes in behavior will correspond to modifications of the defaults.

Problems in designing a testbed that uses simulation

The term "domain" is usually used in AI to describe an intelligent task to be simulated. But it would be a little misleading for us to use the term "domain" in this way, since the conversational tasks we have in mind will involve both an object domain and a meta-domain. That is, there will be reasoning about some nonlinguistic subject, as well as the reasoning involved in generating and interpreting discourse. In choosing a discourse task, we will need to select both a world domain and a discourse genre or discourse domain. We will call the combined task a *discourse task.*

General considerations

Need for a specific world task. In general, natural language processing systems perform well only in the context of a specific world task. To develop a successful discourse simulation, we will certainly need to restrict the world.

Focus on reasoning methods. We need a discourse task that is complex enough to elicit interesting discourse phenomena but not so complicated that simulation is unrealistic. Is there such a task? Can we choose the task to allow us to focus on the discourse phenomena in which we are interested? Probably we can begin to explore these issues in detail only by experimenting with some simulations. If this is right, then the idea of selecting a specific discourse-related phenomenon (like anaphora, or demonstrative reference) will not in itself be a very useful guide to simulation design. An approach that takes into account the reasoning mechanisms on which one wishes to focus looks more promising.

Discourse reasoning tasks divide at once into generation and interpretation tasks. A simulation combining both these elements should use mechanisms that match. It would be a mismatch, for instance, to use an ordinary expert system for generation and a Bayesian network for interpretation, since only one of these mechanisms learns. It is not unusual to think of conversation as a coordination problem between the speaker and the hearer,[2] but it would be unreasonable to make the hearer responsible for all the learning required for solving this problem. Also, the reasoning mechanisms involved in discourse should fit together into the same cognitive architecture, or at least should use styles of reasoning that broadly match each other. In this sense, it would be a mismatch to use SOAR for generation and a Bayesian network for interpretation; though both mechanisms learn, they rely on very different reasoning styles.

If we wish to use the reasoning mechanisms that we have chosen for discourse tasks for world reasoning,[3] decisions about the discourse domain may affect the choice of a world domain. For instance, if plan recognition is an important element of discourse interpretation, it is natural to choose a world domain that also involves planning and plan recognition, and to assume that the same planning and recognition mechanisms used in the world domain are used in the discourse domain.

Measuring the quality of performance. Many of the hypotheses that we would want to test with such an ar-

[1] See (Lascarides & Oberlander 1991), (Lascarides & Asher 1991), (Lascarides & Asher 1993a), and (Lascarides & Asher 1993b).

[2] See (Lewis 1969), (Clark 1992).

[3] The decision is not just a matter of economy. In discourse, we have to address the interface between linguistic and domain reasoning. It would be costly to postulate that special-purpose processing mechanisms that are unrelated to general-purpose forms of reasoning are involved in linguistic tasks. Several traditions in pragmatics (the ordinary language tradition, for instance, and most of the work in AI) explicitly deny this assumption.

chitecture relate other variables to the quality of the resulting discourse. Because of this, the discourse simulation has to provide a way of estimating the quality of performance of a conversational task. And the design should permit at least some variations in outcome to be traced to the conversational strategies of the agents.

The study reported in (Walker 1993) is a good example of this. Walker's broad hypothesis was that redundancy in conversations serves as a way of compensating for memory limitations, with the consequence that—at least, within certain limits—redundancy would be more useful as the memory capacity of conversational agents is restricted. She designed a world domain task in which discussants locate furniture in rooms; a score representing the sum of the values of the furniture items is used as a measure of "usefulness".

Suitability for human subjects. Some hypotheses about conversational strategies will probably have to do with the "naturalness" of the resulting conversation. For instance, (Clark 1992, Chapter 3) suggests that there is relationship between the accuracy of the speaker's model of the hearer's notion of salience and the naturalness of generated referring expressions.

There may well be no way to convert hypotheses like these into quantifiable relationships, or even to come to grips objectively with the general notion of naturalness that is involved here. But we can provide qualitative tests for such hypotheses by choosing a task for simulation that can also be performed with human subjects. It is a great advantage of an environment in which both generation and interpretation are simulated that it can be used in several modes: (1) as a way of producing simulated conversations under a wide variety of conditions, (2) as an agent capable of carrying out the discourse task with a human, and (3) as an environment for performing discourse protocols for human subjects.

This provides several assessments of naturalness. In the example of referring expressions, we could simply compare the referring expressions generated by a system exhibiting systematic misconceptions about salience, with the referring expressions produced by humans in the environment. Here, the judgments about naturalness would have to be subjective. A more objective sort of test would have simulated agents interact with human subjects, who are allowed the option of calling for clarification of referring expressions. The number of human calls for clarification provides a measure of referential naturalness, which can then be compared with the sophistication of the agent's model of salience.

Exploring a specific choice of discourse tasks

The number of potential alternative discourse tasks is very large, and we have not yet found a good way to explore them systematically. So at this point we will shift to the particular, and explain the choices that we ourselves have made in specifying the simulation environment we plan to build.

Focus on coordination. We began by taking our own methodological advice—our choice of discourse task was guided to a large extent by a focus on certain theoretical issues and processing mechanisms. In particular, we wish to concentrate on the theoretical issue of *coordination in conversation.* Coordination is theoretically central for discourse (we believe that it lies at the heart of Grice's "cooperative principle"), and an architecture that incorporates generation and interpretation can hardly avoid taking the interrelationships of the two into account.

Modeling mutual belief. Like many discourse theoreticians (especially Clark),[4] we think of mutual belief as a crucial component of coordination. We assume that planning and plan recognition are the central reasoning mechanisms involved in discourse generation, and interpretation respectively. Further, we have decided to work with abductive models of plan recognition.[5]

In part, our decision to select these reasoning mechanisms was guided by practical considerations; we are collaborating with Johanna Moore's research group in generation at the University of Pittsburgh, and with Jerry Hobbs' group at SRI. But the importance of planning and plan recognition is generally accepted, at least in computational circles. The selection of planning and plan recognition suggests a general hypothesis—that coordination to a large extent is a matter of the attunement of these two mechanisms to each other. That is, the extent to which discourse is well coordinated will depend on the extent to which the planner is well adapted to the plan recognizer, and the extent to which the plan recognizer is well adapted to the planner.

Coordination is not easily modeled in terms of an analysis of speech acts based on conversants' beliefs about each others' beliefs; this is related to the fact that mutual belief is not definable in terms of the beliefs of group members. We have selected a different model, inspired by work of Stalnaker's,[6] and will split the beliefs maintained by each conversational agent into two parts, representing (1) private beliefs, and (2) the agent's view of what is "common ground" in the conversation.

A way to measure coordination. With this separation in the attitudes of each agent between private beliefs and common ground, we now have a way to measure

[4]There also is a connection to the theoretical ideas in (Thomason 1990a) and (Thomason 1990b).

[5]See (Hobbs et al. 1994).

[6]See (Stalnaker 1978); this idea also fits well with the ideas of (Clark 1992).

the coordination that has been achieved in a simulated conversation; we can simply compare the two agents' views of what has been established in the conversation. To take an example from our domain, suppose that there are two green chairs, one in the living room and one in the bedroom, and that agent A says "The green chair costs $150," meaning the green chair in the living room. Agent A will then add something like

```
Cost(Chair-6, 150)
```

to its record of the common ground. Agent B, misunderstanding the utterance, adds

```
Cost(Chair-13, 150)
```

to its view of the common ground. In this example, then, the discrepancy in the two views of the common ground will record a miscoordination.

The discourse task. Our discourse task is an adaptation of the task domain in (Walker 1993), where two agents collaborate on furnishing a number of rooms. The differences between our research goals and Walker's require substantial modifications of her task. But the choice of a collaborative activity calling for negotiation and exchange of knowledge in setting and achieving common goals suits our interests well, and we believe that there may be some general advantage to the research community in some standardization of discourse tasks. For example, a common collaborative design task would be a good standard for studying issues in coordination and negotiation. By doing so, competing hypotheses could be comparably evaluated and refined. However, a different type of task would be necessary for other types of discourse (e.g. adversarial discourse such as courtroom cross-examinations, and tutorial discourse).

In the original task defined in (Walker 1993), the agents know the values of all the furniture that can be allocated; each agent is given a personal allotment from this set. The main goal for all the variants of the task is to pick the 8 highest valued pieces of furniture that can be remembered.

The modifications in our version of the discourse task reflect the differences in the theoretical issues we wish to explore.

1. At least initially, we are not investigating issues related to limited recall. We assume that the agents have perfect recall of the domain information accessible to them at the beginning of the task. (I.e. they can refer to lists of furniture and prices.)
2. We depart from a rigid script for exchanges, since (among other things) we want to investigate speech act recognition as a collaborative activity.
3. The collaborative planning activity requires at least the availability of *bids* (or proposals) and acceptances or rejection of these bids. To promote other sorts of speech acts (and to provide enough difference in agents' perspective to generate interesting dialog), we give agents privileged access to some aspects of the world domain. In particular, the available furniture is parcelled out among various "warehouses;" warehouses are assigned to agents. Agents only know the inventory of their personal warehouses (and this fact is mutually known by the agents). Agents are also assigned budgets; each agent only knows his own budget. These knowledge limitations will introduce domain motives for questions; and, of course, questions call for assertions as answers.
4. Collaboration is ensured by insisting that decisions about furniture placement can only be made if approved by both agents. We have devised a scoring system that tries to quantify the success with which groups maximize the "furnishing value" of their joint budgets by placing appropriate furniture in their rooms. In more complex tasks, we may assume that certain rooms are assigned to specific agents, who have sole responsibility over furniture placement, unless they need an item from a partner's warehouse. (We believe this change will enforce more complex forms of collaboration.)
5. The agents must collaborate during the means-end reasoning to arrive at options for deliberation (assuming the intelligent resource-bounded machine architecture (IRMA) (Pollack 1992) upon which the simulation in (Walker 1993) is based.)

Illustrating additional hypotheses

Many of the hypotheses we wish to explore have to do with the nature of planning and plan recognition in natural conversation. We close the discussion of our proposed simulation testbed by mentioning one such example. Assuming a SharedPlan model of the domain plan,[7] we can investigate how different discourse strategies depend on whether the currently salient plan is shared or individual. Here is one such hypothesis: when another agent's proposal conflicts with a particular type of plan, it will trigger a particular discourse action as follows.

- A shared plan triggers a coordination check.
- The hearer's individual plan triggers a negotiation.
- The model of the other agent's individual plan triggers a question.

With a simulation we could determine whether such a dialog strategy results in improved or degraded performance.

[7] See (Grosz & Kraus 1993).

Our Progress in Building a Simulation Testbed

We expect to have the first version of the simulation testbed completed by the end of the summer. So far we have built a dialog collection environment for the discourse task we described earlier. The collection environment allows us to record the conversation between the human subjects as well as the degree to which the subjects are coordinated throughout the conversation. We have collected 18 trials using 6 subject pairs with each trial averaging approximately 100 clauses. We have begun to analyze this data for plan structure and are currently investigating building a task planner for the simulation testbed based on the preliminary results of our analysis. We are also investigating generation in an abductive framework to complement abductive interpretation.

Acknowledgments

This material is based on work supported by the National Science Foundation under Grant No. IRI–9314961. We wish to acknowledge the contributions of fellow project members Megan Moser, Johanna Moore, and Jerry Hobbs. We also thank Marilyn Walker for her insights and for giving us access to the DesignWorld Simulation environment.

References

Herbert Clark. *Arenas of language use.* University of Chicago Press, Chicago, 1992.

Barbara Fox. "Interactional Reconstruction in Real-Time Language Processing." *Cognitive Science 11* (1987), pp. 365-387.

Barbara Grosz and Sarit Kraus. "Collaborative plans for group activities." Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, 1993, Morgan Kaufmann, Los Altos, 1993.

Jerry Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. "Interpretation as abduction." *Artificial Intelligence 63* (1993), pp. 69–142.

Pamela Jordan and Megan Moser. "Multi-level Coordination in Computer-mediated Conversation." *GLS 1995: Developments in Discourse Analysis,* Georgetown University, 1995. Forthcoming.

Alex Lascarides and Jon Oberlander. "Temporal coherence and defeasible knowledge." *Proceedings of the workshop on discourse coherence,* Edinburgh, 1991.

Alex Lascarides and Nicholas Asher. "Discourse relations and common sense entailment." *Proceedings of the 29th Meeting of the Association for Computational Linguistics,* 1991, pp. 55–63.

Alex Lascarides and Nicholas Asher. "Temporal interpretation, discourse relations and common sense entailment." *Linguistics and philosophy 16* (1993), pp. 437–494.

Alex Lascarides and Nicholas Asher. "Lexical disambiguation in a discourse context." Unpublished MS, 1993.

David Lewis. *Convention.* Harvard University Press, Cambridge, Massachusetts, 1969.

Allen Newell. "You can't play 20 questions with nature and win: projective comments on the papers at the symposium." In W.G. Chase, ed., *Visual Information Processing,* Academic Press, New York, 1973, pp. 283–308.

Allen Newell. *Unified Theories of Cognition.* Harvard University Press, Cambridge, Massachusetts, 1990.

Martha Pollack. "The Uses of Plans." *Artificial Intelligence 57* (1992), pp. 43–68.

Robert Stalnaker. "Assertion." In P. Cole, ed., *Syntax and Semantics 9: Pragmatics,* Academic Press, New York, 1978, pp. 315–332.

Richmond Thomason. "Accommodation, meaning, and implicature: interdisciplinary foundations for pragmatics." In P. Cohen, J. Morgan, and M. Pollack, eds., *Intentions in communication.* MIT Press, Cambridge, Massachusetts, 1990, pp. 325–363.

Richmond Thomason. "Propagating epistemic coordination through mutual defaults I." In R. Parikh, ed., *Proceedings of the Third Conference on Theoretical Aspects of Reasoning about Knowledge.* Morgan Kaufmann, 1990, pp. 29–39.

Marilyn Walker. *Informational redundancy and resource bounds in dialog.* Ph.D. Dissertation, University of Pennsylvania, Philadelphia, 1993.