Running head: DISCOURSE AND CONVERSATION

Computational Modeling of Discourse and Conversation

Arthur C. Graesser, Danielle S. McNamara, and Vasile Rus

University of Memphis

Send correspondence to:

Art Graesser
Psychology Department
202 Psychology Building
University of Memphis
Memphis, TN, 38152-3230
901-678-2742
901-678-2579 (fax)
a-graesser@memphis.edu

Draft of chapter to the *Cambridge Handbook of Psycholinguistics*, edited by Michael Spivey, Marc Joanisse, and Ken McRae.

Sentences and spoken utterances are nearly always expressed in the context of a text or conversation. It is therefore important to understand the mechanisms that drive the comprehension and production of connected discourse. In this chapter we will use the term *discourse* as a covering term for text, monologues, dialogues, and multiparty conversations. Our goal is to understand the structures, representations, strategies, and processes that underlie the comprehension and production of discourse. There is a field, called *discourse processes,* that is devoted to scientific investigations of these mechanisms. It has its own journal (*Discourse Processes*), affiliated society (Society for Text and Discourse), and its own *Handbook of Discourse Processes* (Graesser, Gernsbacher, & Goldman, 2003).

This chapter focuses on computational models of discourse processing. There are two senses of computation, both relevant to this chapter. The first sense refers to the architectures and algorithms in models of human discourse processing. The second sense refers to computer implementations of components of these psychological models. Some psychological components can be programmed in computers, whereas others are currently beyond the immediate technological horizon. We hope to clarify the boundaries of what is technically feasible on computers, knowing full well that the boundaries change yearly.

There are many types of discourse, or what some researchers call *genre* (category in French), *registers,* or simply discourse categories. Discourse typologies vary in grain size and theoretical slant, with some researchers viewing the discourse landscape as a set of fuzzy categories, others a structured ontological hierarchy, and yet others viewing it as a multidimensional space (Biber, 1988). There are prototypical discourse categories in

the American culture, such as folktales, scientific journal articles, jokes told in stand-up comedy, and one-on-one tutoring. These four examples would funnel into more superordinate classes that might be labeled as narrative, exposition, monologue, and dialogue, respectively. Of course, there will always be blends, hybrids, and borderline cases, such as a faculty member chatting with a student about a funny story involving an experiment that failed. Whether there is a set of prototypical discourse categories is a lively topic of debate.

This chapter will concentrate on two forms of discourse: text comprehension and two-party dialogue. We acknowledge that there are other important forms of discourse, such as text production, comprehension and production of spoken monologues, and multi-party conversations with three or more participants. However, most research has been conducted on text comprehension and two-party dialogue.

Computational Models of Text Comprehension

Levels of representation

Multiple levels of representation are constructed when a text is comprehended. Five of these levels are the surface code, the propositional textbase, the situation model, the text genre, and the pragmatic communicative context (van Dijk & Kintsch, 1983). Suppose, for illustration, that the following excerpt about a corporation was read in a newspaper:

> When the board met on Friday, they discovered they were bankrupt. They needed to take some action, so they fired the president.

The *surface code* preserves the exact wording and syntax of the sentences. The *textbase* contains explicit propositions in the text in a stripped down, logical form that preserves

the meaning but not the surface code. The first sentence would have the following propositions, a theoretical construct that will be elaborated shortly:

PROP 1: meet (board, TIME = Friday)

PROP 2: discover (board, PROP 3)

PROP 3: bankrupt (corporation)

PROP 4: when (PROP 1, PROP 2)

The *situation model* (sometimes called mental model) is the referential content or microworld that the text is describing. This would include the people, objects, spatial setting, actions, events, plans, thoughts and emotions of people and other referential content in a news story, as well as the world knowledge recruited to interpret this contextually specific content. The *text genre* is the type of discourse, in this case a news story about a corporation. The pragmatic *communicative context* is the implicit dialogue between the author (story writer, editor) and the reader (a public citizen). The story was presumably written to convey some point to the reader, such as the particular corporation is on the brink of collapse. The public, of course, loves disaster stories.

Discourse context and world knowledge are extremely important in guiding the construction of these levels of representation. The referents and propositional content of the textbase would not be composed correctly if one relied on the local sentence context. For example, the *they* in the phrase *they were bankrupt* refers to the corporation rather than the board, yet models that assign referents to pronoun anaphors on the basis of sentence syntax would make just the opposite assignment. The *president* refers to the president of the board, not the president of the United States, even though the U.S. president is shared knowledge among U.S. citizens and the modifying determiner is

definite (*the*). These assignments are rather subtle and require a fine-grained analysis of context and world knowledge. This subtlety and complexity become more salient as soon as a researcher tries to get computers to perform these computations.

Researchers are not entirely in agreement that the explicit discourse can be segmented into a structured set of propositions. One problem with the construct of a proposition is that researchers from different fields (i.e., artificial intelligence, logic, linguistics, and psychology) do not entirely agree on the format and formal constraints of propositions. In the field of discourse psychology, a proposition refers to a state, event, or action that may or may not have a truth value with respect to the referential situation model; this contrasts with propositional calculus theories of traditional logic where a truth value must be assigned and the meaning of the proposition is unimportant. In psychology, each proposition contains a predicate (e.g., main verb, adjective, connective) and one or more arguments (e.g., nouns, embedded propositions). In most notational systems, the arguments are placed within the parentheses, whereas the predicates are outside of the parentheses. Each argument has a functional role, such as agent, patient, object, time, or location, although the theoretical set of functional roles differs somewhat among the fields of psychology, linguistics, artificial intelligence, and computational linguistics (Allen, 1995; Kintsch, 1998; Van Dijk & Kintsch, 1983). Discourse psychologists sometimes ignore the role of quantifiers (one, some, all) when propositional representations are constructed, whereas quantifiers are explicitly captured in the predicate calculus representations (called first order logic) of artificial intelligence, computational linguistics, and formal logic. There are yet other differences among fields

that pertain to structural composition and the epistemological status of propositions (e.g., facts, beliefs, wants), far too many to enumerate here.

There would be tremendous advantages to having a computational model that translates the language of discourse into the logical forms in logic or AI (such as first-order predicate calculus) or the deep structures in standard linguistics. One advantage of a logical form is that well established computational procedures can execute theorem proving and inference generation in an elegant manner. Another advantage is that discourse structures and world knowledge structures would have a uniform representation that could be systematically aligned, compared, and integrated. Unfortunately, there are two serious challenges about this neat and tidy picture. The first challenge is that researchers in AI and computational linguistics have not been able to develop a computer program that can reliably translate discourse constituents into a logical form or deep structure representations, even in large-scale evaluations that aspire to such a goal (Rus, 2004). The vast majority of today's syntactic parsers, such as Apple Pie (Sekine & Grishman, 1995) and the Charniak parser (2000) construct tree structures that capture surface structure composition rather than deep structures or logical forms. The second challenge is that it may not be necessary, psychologically, to fuss with the construction of a propositional textbase. Instead, the words and other linguistic signals in the surface code might provide a direct route to the situation model and other meaning representations (Zwaan & Radvansky, 1998). Of course, those who dismiss the construct of a proposition need to have a principled way of specifying meaning representations and how they systematically get constructed. As yet, an alternative to propositions has not

been developed, apart from a few proposals that can handle only a small corpus of examples.

World knowledge

A computational model of discourse comprehension must make some commitments to the representation of world knowledge. World knowledge is needed to guide the interpretation of explicit information and also to furnish plausible inferences (Graesser, Singer, & Trabasso, 1994; Kintsch, 1998). Three theoretical frameworks for handling world knowledge are conceptual graph structures (Lehmann, 1992; Schank & Reisbeck, 1982; Sowa, 1983), high dimensional semantic spaces (Landauer, Foltz, & Laham, 1998; Landauer, McNamara, Simon, & Kintsch, in press), and embodied representations (Glenberg, 1997). Conceptual graph structures (CGS's) contain a set of nodes (referring to noun-like concepts or propositions) that are interrelated by labeled, directed arcs, such as Is-a, Has-as-parts, Cause, and Reason. The referent of a CGS may include a family of related concepts (such a semantic network for animals), a package of nodes that capture a specific experience (e.g., a previous experience or text that is read), or a generic package of knowledge, such as a script on how people dine in restaurants, a stereotype about professors, or a schema about corporate bankruptcy (Schank & Reisbeck, 1982). During the course of comprehending a particular text, a family of these background CGS's get activated and guide the interpretation of sentences and the generation of inferences. The CGS approach was the dominant approach to representing world knowledge between the late 1970's to late 1990's.

One salient limitation of the CGS's is that the researcher has to construct the content and relations by hand. Progress on automatic construction of these structures

through machine learning algorithms and theoretical formal systems has not scaled up to handling a large corpus of texts, although there were some notable successes in inducing noun concept taxonomies in semantic networks (Hearst, 1992; Stevenson, 2002) and case hierarchies in case-based reasoning (Veloso & Carbonell, 1993).

In the mid-1990's the zeitgeist shifted from handcrafted structures to high dimensional conceptual spaces that accommodate the constraints of a large corpus of texts. Notable examples of statistical, corpus-based approaches to analyzing the world knowledge that underlies discourse are the Hyperspace Analog to Language (Burgess, Livesay, & Lund, 1998), Latent Semantic Analysis (Kintsch, 1998; Landauer, Foltz, & Laham, 1998; Landauder et al. in press), and the Linguistic Inquiry Word Count (Pennebaker & Francis, 1999). LSA uses a statistical method called "singular value decomposition" (SVD) to reduce a large Word-by-Document co-occurrence matrix to approximately 100-500 functional dimensions. The Word-by-Document co-occurrence matrix is simply a record of the number of times word W_i occurs in document D_j. A document may be defined as a sentence, paragraph, or section of an article. Each word, sentence, or text ends up being a weighted vector on the K dimensions. The "match" (i.e., similarity in meaning, conceptual relatedness) between two unordered bags of words (single words, sentences, or texts) is computed as a geometric cosine between the two vectors, with values ranging from -1 to 1. LSA-based technology is currently being used within a number of applications, such as essay graders that grade essays as reliably as experts in English composition and automated tutors that give feedback equivalent to human tutors (see chapters in Landauer et al., in press).

One limitation in both CGS's and LSA is that they gloss over many of the fine details of perceptual experiences and motor activity. CGS's are symbolic and LSA is a statistical representation. In contrast, the embodied framework grounds discourse in sensori-motor experience and constraints of the body as the body interacts with a particular world (Glenberg, 1997; Roy, 2005). It should be acknowledged that some structural theories ground the symbolic nodes (concepts, propositions) in perception and action, and there are LSA-based models that are capable of representing sensori-motor procedures with a suitable corpus. However, an embodied framework is arguably needed to go the full distance in handling references to perception, action, deixis, and experiences that ground symbols. Unfortunately, no one has built a model on a computer that is capable of generating fully embodied representations from naturalistic text and of inducing embodied representations of world knowledge from experiences. There are robotic systems that ground words and simple spoken utterances in perception and action (Roy, 2005), but there are no systems that take text input and automatically produce a representation that is even close to an embodied representation. This is one challenge for future research.

Cohesion and coherence

Sentences and clauses in connected discourse need to be coherently related in order to convey the desired message to the reader. A distinction is often made between *cohesion* and *coherence* (Graesser, McNamara, Louwerse, & Cai, 2004; van Dijk & Kintsch, 1983). Cohesion is an objective property of the explicit text. Explicit words, phrases, sentences, and linguistic features guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher

level global units (e.g., topics, themes). Coherence refers to the quality of the mental representation constructed by the comprehender. Cohesive devices cue the reader how to construct a coherent representation in the mind; how and whether this happens, however, depends on the skills and knowledge the reader brings to the situation (McNamara & Kintsch, 1996). For example, if the reader has adequate world knowledge about the subject matter or if there are adequate linguistic and discourse cues, then the reader is likely to form a coherent mental representation of the text. Readers follow an underlying pragmatic assumption that texts are coherent and expend effort to construct coherent representations while reading well constructed texts. However, if text is very poorly composed, their efforts fail so they give up trying and attribute problems to either the text or their own deficits in world knowledge.

Coh-Metrix is a computer tool available on the web that analyzes texts on multiple levels of cohesion, as well as other levels of language (http://cohmetrix.memphis.edu, Graesser, McNamara, Louwerse, & Cai, 2004). Coh-Metrix has the potential to replace standard readability formulas, such as Flesch-Kincaid Grade Level (Klare, 1974-1975), which rely exclusively on word length and sentence length to scale texts on readability. The user of Coh-Metrix enters a text into the web site and it prints out measures of the text on 44 metrics that span different levels of discourse and language. Coh-Metrix 1.3 has measures in the following categories: (1) co-referential cohesion, such as nouns referring to other nouns and phrases; (2) causal cohesion; (3) density of different categories of connectives and logical operators; (4) LSA-based conceptual cohesion; (5) type-token ratio; (6) readability measures; (7) word frequency measures; (8) density of words in different parts of speech; (9) other word

characteristics, such as concreteness, polysemy, and age of acquisition; (10) density of noun-phrases; and (11) syntactic complexity. Coh-Metrix integrates lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters, LSA, and other components that have been developed in the field of computational linguistics (Allen, 1995; Jurafsky & Martin, 2000). For example, Coh-Metrix incorporates several lexicons, including CELEX (Baayen, Piepenbrock, & Van Rijn, 1993), WordNet (Fellbaum, 1998), and the MRC Psycholinguistic Database (Coltheart, 1981). These lexicons allow us to measure each word on number of syllables, abstractness, imagery, ambiguity, frequency of usage, age of acquisition, number of senses (meanings), and dozens of other dimensions. There is a part-of-speech "tagger" (Brill, 1995) that assigns each word to one of 56 syntactic classes; it uses context to assign the most likely class when a word can be assigned more than one part of speech. There is a syntactic parser that assigns syntactic tree structures to sentences and measures them on syntactic complexity (Sekine & Grishman, 1995). The LSA module measures the conceptual similarity between sentences, paragraphs, and texts on the basis of world knowledge.

Coh-Metrix 2.0 has been expanded to incorporate more levels of cohesion in discourse. It computes the referents of pronouns on the basis of syntactic rules, semantic fit, and discourse pragmatics by some existing algorithms proposed by Mitkov (1998) and Lappin and Lease (1994). It also segregates different dimensions of the situation model, including those of agency, temporal, spatial, causal, intentional, and logical cohesion. As in several models in discourse psychology, these dimensions were included in Zwaan and Radvansky's (1998) event indexing model; their review of the psychological literature confirmed that incoming sentences take more time to read to the

extent there are coherence gaps in agency, temporality, spatiality, intentionality (i.e., goals, plans, actions of agents), and causality.

To compute intentionality, Coh-Metrix has an algorithm that identifies actions and goals by combining syntactic information with information from the WordNet database (Fellbaum, 1998). For example, intentional actions and goals have main verbs that are either causal or intentional (as defined by a cluster of lexicographical categories in WordNet) and *animate* or *human* subject nouns (e.g., in *the girl bought a car*, the verb *buy* is intentional and the subject noun *girl* is human). A text is cohesive on the intentional dimension to the extent that there are more intentional linguistic particles that link actions and goals, such as conjunctions and other forms of connectives (e.g., *in order to, so that*).

Coh-Metrix also has measures on structural cohesion, including syntactic similarity of sentences, ease of identifying topic sentences, genre uniformity, document headings, and given-new information contrasts. One of the most influential analyses of genre has been that of Biber (1988), who used factor analysis to classify a large corpus of texts on the basis of 67 features of language and discourse. Coh-Metrix 2.0 has automated 62 out of 67 of these features so it can compute the extent to which a text fits different genres (such as narrative, science, versus history texts). Associated with each genre is a diagnostic set of connectives, discourse markers, and other signaling devices. Discriminant function analyses identify the features that diagnostically predict whether text *T* is in genre/class *G*. Texts can thereby be scaled on global genre cohesion in two ways. First, a text has higher genre cohesion when it cleanly fits into one prototypical genre/class *G* (as measured by an inverse of the classification entropy score). Second,

there is higher global cohesion when there is a higher density of diagnostic features associated with the dominant genre/class *G*.

Another analysis of structure contrasts *new* from *given* information by segregating constituents that are introduced for the first time in the text from references to previous text constituents and from information that is in the common ground (shared knowledge) of speech participants (Prince, 1981). Whereas previous analytical treatments of the given-new distinction have been compositional and symbolic, Coh-Metrix 2.0 uses an LSA algorithm to segregate *new* versus *given* information as sentences are comprehended, one by one.

Computational models of text comprehension in humans

Discourse psychologists have developed a number of models that simulate how humans comprehend text. Among these are the Collaborative Action-based Production System (CAPS) Reader model (Just & Carpenter, 1992), the Construction-Integration model (Kintsch, 1998), the constructivist model (Graesser, Singer, & Trabasso, 1994), and the landscape model (Van den Broek, Virtue, Everson, Tzeng, & Sung, 2002). The architectures of these models go beyond simple finite state automata that have a small finite set of states (categories) and a small set of transition matrices (one per process) that specify the likelihood that a theoretical entity will change states. Rather, they are complex dynamical models with a very large or infinite state space that can evolve in complex and sometimes chaotic trajectories. It is impossible to sufficiently capture these models with a set of linear equations or with a set of simple rules. This subsection will describe the CAPS/Reader and Construction-Integration (CI) model.

Just and Carpenter's (1992) CAPS/Reader model directs comprehension with a large set of production rules. The CAPS/Reader model is a hybrid between a production system and a connectionist computational architecture. Each of the production rules (a) scan explicit text input, (b) govern the operation of working memory, (c) change activation values of information in working memory and long-term memory, and (d) perform other cognitive or behavioral actions. Production rules have an "If <state>, then <action>" form, but these rules are probabilistic, with activation values and thresholds, rather than being brittle. If the contents of working memory has some state S that is activated to a degree that meets or exceeds some threshold T, then action A is executed by spreading activation to one or more other information units in working memory, long-term memory, or response output. A state slot may be arbitrarily complex, often consisting of several sub-states that capture a pattern of language or discourse. For example, consider a possible rule that would identify intentional actions: If the subject noun is animate and the main verb is causal or intentional, then activation is spread to the proposition category of intentional action. The proposition would be classified as intentional only probabilistically, because other activated production rules may spread activation in a fashion that does not converge on the category of intentional action. All of the production rules are evaluated in parallel within in each cycle of the production system, and multiple rules may get activated within each cycle. The researcher can therefore trace the activation of information units (nodes) in the text, working memory, and long-term memory as a function of the cycles of production rules that get activated. Just and Carpenter have reported these profiles of nodal activation can predict patterns of

reading times for individual words, eye tracking behavior, and memory for text constituents.

Kintsch's (1998) CI model directs comprehension with a connectionist network. As text is read, sentence by sentence (or alternatively, clauses by clause), a set of word concept and proposition nodes are activated (constructed). Some nodes match constituents in the explicit text whereas others are activated inferentially by world knowledge. The activation of each node fluctuates systematically during the course of comprehension, sentence by sentence. When any given sentence S (or clause) is comprehended, the set of activated nodes include (a) N explicit and inference nodes affiliated with S and (b) M nodes that are held over in working memory from the previous sentence S-1 by virtue of meeting some threshold of activation. As a consequence, there are N+M nodes to reckon with while comprehending sentence S. These N+M nodes are fully connected to each other in a weight space. The set of weights in the resulting (N+M) by (N+M) *connectivity matrix* specifies the extent to which each node activates or inhibits the activation of each of the other N+M nodes. The values of the weights in the connectivity matrix are theoretically motivated by multiple levels of language and discourse. For example, if two word nodes (A and B) are closely related in a syntactic parse, they would have a high positive weight, whereas if two propositions contradict each other, they would have a high negative weight.

The dynamic process of comprehending sentence S has a two stage process, namely construction and integration. During construction, the N+M nodes are activated and there is an initial activation vector for the set of nodes (a_1, a_2, …a_{N+M}). The connectivity matrix then operates on this initial node activation vector in multiple

activation cycles until there is a settling of the node activations to a new final stable activation profile for the N+M nodes. At that point, integration of the nodes has been achieved. Mathematically, this is accomplished by the initial activation vector being multiplied by the same connectivity matrix in multiple iterations until the N+M output vector of two successive interactions shows extremely small differences (signifying a stable settling of the integration phase). Sentences that are more difficult to comprehend would presumably require more cycles to settle. These dynamic processes have testable implications for psychological data. Reading times should be correlated with the number of cycles during integration. Recall of a node should be correlated with the number of relevant sentences and cycles of activation. Inferences should be encoded to the extent that they are activated and survive the integration phase. Kintsch (1998) summarizes substantial empirical evidence that supports these and other predictions from the CI model.

One weakness of the CI model has concerned the connectivity matrix, which captures the core of the language and discourse constraints. In the early days of CI modeling, the researchers had to compose the weights in the connectivity matrix by hand. This approach falls prey to the criticism that the researchers finagled the weights to fit the data in an ad hoc fashion. The obvious exit out of this loop is to generate the weights in a principled fashion computationally, ideally by a computer. The field of computational linguistics is close to achieving such a goal. Kintsch (1998) has used LSA to automatically activate concepts (near neighbors) from long-term memory that are associated with explicit words and to generate weights that connect the N+M nodes. Syntactic parsers can be used to compute weights by virtue of structural proximity. One

technical limitation that researchers are facing is that there is no reliable mechanism for translating language to propositions, an important functional unit in the CI model. One of the frontiers of the CI model is to identify principled automated mechanisms for generating the weights in the connectivity matrices and initial activation values of nodes during sentence comprehension.

Two-party Dialogue

Discourse analysts have identified dialogue patterns in different types (registers) of two-party dialogue. Some patterns are context-free in the sense that they occur in most conversational registers. Context-free patterns include the adjacency pairs in two party dialogue identified by Schegloff and Sachs (1973), such as [question → answer] and [offer → {acceptance/refusal}]. Another ubiquitous pattern is an embedded counter-clarification question (Schober & Conrad, 1997), as illustrated below, in the context of a survey interview.

Person 1 (survey interviewer): How many pets are in your home?

Person 2 (interviewee): Should I include fish?

Person 1: Only include mammals and birds.

Person 2: Okay, I have 4 pets.

The embedded question is of course constrained by the knowledge state of person 2, namely the uncertainty about what constitutes a pet. Another dialogue pattern that is frequent in classrooms is the [Initiate → Response -→ Evaluation] sequence (Sinclair & Coulthard, 1975), or more specifically the [Question → Answer → Feedback] sequence.

Teacher: What's 6 X 9? (Initiation, Question)

Student: 54 (Response, Answer)

Teach: Very good. (Evaluation, Feedback)

In tutorial dialogue, this sequence is expanded into the 5-step tutoring frame introduced by Graesser and Clark (1994).

Teacher: Why is it warmer in the summer than the winter here? (Question)

Student: The earth is closer to the sun? (Answer)

Teacher: I don't think so. (Short Feedback)

Teacher & Student: <Collaborative multi-turn exchange to improve answer>

Teacher: Do you understand? (Comprehension gauging question)

Student: Yeah.

One reason why tutoring is better than classroom instruction is attributable to step 4, where the student and teacher have a collaborative exchange that scaffolds explanatory reasoning

Representing regularities in dialogue

Discourse analysts have documented discourse patterns such as these that occur in different discourse registers. In order to make some progress, they typically segment the conversations into speech act units and assign each unit to a speech act category. For example, D'Andrade and Wish (1985) have developed a system that is both theoretically grounded and that trained judges can reliably use. Their categories include: question (Q), reply to question (RQ), assertion (A), directive (D), indirect directive (ID), expressive evaluation (E), short verbal response (R, including back channel feedback , e.g., *uh huh*), and nonverbal response (N, such as head nod). There has been an ongoing effort to improve the categorization (tagging) of dialogue acts in the Discourse Resource Initiative (Core, Ishizaki, Moore, Nakatani, et al., 1999). Once the speech acts are tagged,

sequences of these categories can be analyzed to test theoretical claims about discourse or to discover new patterns through machine learning inductive procedures.

It is convenient to represent the dialogue regularities by either a set of production rules, a finite state automata, or a state transition network. The number of constraints in the condition slots of the production rules is a rough index of the extent to which the rules are context sensitive versus context free. Some researchers believe there are a large number of conceptual and pragmatic constraints that must be in place before a speech act can be classified in a particular category. A smooth dialogue with a conversational partner may be unlikely without the system having internal representations of the knowledge, beliefs, goals, intentions, plans, norms, values, and other cognitive representations of speech participants (Allen, 1995). Consider, for example, a symbolic structure of the speech act that is functionally a **request**.

REQUEST (S, H, ACT)

[Meaning the speaker requests the hearer to perform an action]

Constraints: **Speaker(S) Λ Hearer (H) Λ ACT(A) Λ**

H is agent of ACT

Precondition: **WANT (S, ACT (H))**

[Meaning the speaker wants the hearer to do the action]

Effect: **WANT (H, ACTION (H))**

[Meaning the hearer wants to do the action after the request]

BELIEVE (H,WANT(S (ACT(H)))

[The hearer believes the speaker wants the hearer to do the action]

A speech act allegedly does not function as a request if the speaker does not want action A to occur, and is ineffective if the listener does not realize the speaker wants A.

The process of reliably reconstructing these knowledge states is of course a challenge computationally. Some progress is being made by statistical algorithms that induce patterns from large corpora of speech samples in particular applications, such as dialogues about train or airline schedules, but a general purpose pattern recognizer of knowledge, goals, and plans is not in the immediate technological horizon (Jurafsky & Martin, 2000). There must be a close alignment in the common ground (shared knowledge) between the two participants in the dialogue before there can be a reliable recognition of the other's knowledge, goals, and plans, but the process of achieving such an alignment is under debate (Clark, 1996; Pickering & Garrod, 2004).

Computer simulation of dialogue

Weizenbaum's (1966) Eliza program was the first conversational system that was reasonably successful, popular, and widely used. Eliza simulated a Rogerian client-centered psychotherapist, as illustrated in the conversation below.

Patient: You are like my father in some ways.

ELIZA: How am I like your father?

Patient: You are very aggressive but I think you don't want me to notice that.

ELIZA: What makes you think I am aggressive?

Patient: You always argue with me.

ELIZA: Why do you think I argue with you?

Like a typical Rogerian therapist, Eliza tried to get the patient to do the talking by asking the patient questions about the patient's verbal contributions. Eliza turned the patient's

assertions into therapist's questions by simple syntactic transformational rules. That is, Eliza detected keywords and word combinations that matched production rules, which in turn generated Eliza's responses. The only intelligence in Eliza was the stimulus-response knowledge captured in production rules that operate on keywords and that perform syntactic transformations. What was so remarkable about Eliza is that 100-200 simple production rules could very often create an illusion of comprehension, even though Eliza had minimal depth and common ground with the user.

Efforts to build conversational systems continued in the 70's and early 80's. Schank and his colleagues built computer models of natural language understanding and rudimentary dialogue about scripted activities (Schank & Reisbeck, 1982). SHRDLU manipulated simple objects in a blocks world in response to a user's command (Winograd, 1972). By the mid-1980's, however, researchers were convinced that the prospect of building a good conversational system was implausible. The chief challenges were (a) the inherent complexities of natural language processing, (b) the unconstrained, open-ended nature of world knowledge, and (c) the lack of research on lengthy threads of connected discourse. This pessimistic picture was arguably premature because there have been a sufficient number of technical advances in the last decade for researchers to revisit the vision of building dialogue systems. The current conversational systems are not perfect, but they go a long way in creating the impression that the system is comprehending the user and responding appropriately.

A plan-based architecture is routinely adopted in current systems with dialogue modeling in computational linguistics, such as TRINDI (Larsson & Traum, 2000) and COLLAGEN (Rich, Sidner, & Lesh, 2001). The TRINDI project assumes the existence

of an information state, that is, a rather detailed record of the current state of the dialogue. The information state is sufficiently detailed at multiple levels of language and planning to make the particular dialogue distinct and to support the successful continuation of the dialogue. The information state approach is general enough to accommodate dialogue systems that range from the simplest finite-state script to the most complex Belief-Desire-Intention (BDI) model. The information state theory of dialogue modeling requires: (1) a description of the components of the information state, (2) formal representations of these components, (3) external dialogue from the human/other which triggers the update of the information state, (4) internal update rules which select dialogue moves and update the information state, and (5) a control strategy for selecting update rules to apply, given a particular information state. The COLLAGEN project (Rich et al., 2001) is very similar but contrasts three kinds of structure: linguistic, intentional, and attentional. Linguistic structure captures the sequence of utterances, whereas intentional structure captures the conversation goals, and the attentional state is the focus of attention on salient elements of the discourse at a particular point. Existing implementations of COLLAGEN are an approximation of an underlying discourse theory of Grosz and Sidner (1986).

Natural language dialogue (NLD) facilities are expected to do a reasonable job in some conversational contexts, but not others. It depends on the subject matter, the knowledge of the learner, the expected depth of comprehension, and the expected sophistication of the dialogue strategies. A NLD facility is progressively more feasible when more of the following conditions are met.

1) An imperfect system is useful.
2) Expected precision of the information is modest.

3) Content is verbal content rather than mathematical.
4) The user has low or modest subject matter knowledge.
5) Idiomatic expressions are rare.
6) The computer doesn't need to construct a novel mental model.
7) The computer anticipates what users will say.
8) There are simple pragmatic ground rules.
9) The computer has many options and a license to redirect the dialogue by changing topics, asking questions, expressing generic dialogue moves (*Uh huh, Anything else?, I don't follow, That's interesting*).

Tutorial dialogue systems

Tutoring environments are feasible NLD applications because they meet most or all of the above nine conditions, particularly when the subject matter is verbal. It is noteworthy that even human tutors are not able to monitor the knowledge of students at a precise fine-grained level because much of what students express is vague, underspecified, ambiguous, fragmentary, and error-ridden (Fox, 1993; Graesser & Person, 1994). There are potential costs if a tutor attempted to do so. For example, it is often more worthwhile for the tutor to help build new correct knowledge than to become bogged down in dissecting and correcting each of the learner's knowledge deficits. Tutors do have an approximate sense of what a student knows and this appears to be sufficient to provide productive dialogue moves that lead to significant learning gains in the student (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Graesser & Person, 1994).

Researchers have developed approximately a half dozen intelligent tutoring systems with dialogue in natural language. These systems help college students generate

cognitive explanations and patterns of knowledge-based reasoning when solving particular problems (Moore, 1995). *AutoTutor* (Graesser, Olney, Haynes, & Chipman, 2005) was developed for introductory computer literacy and Newtonian physics. *Why/Atlas* (VanLehn, Jordan, et al., 2002) also has students learn about conceptual physics by a coach that helps them build explanations. The *Mission Rehearsal* system (Gratch et al., 2002) helps Army personnel interact in a virtual war scenario. *iSTART* helps students learn reading strategies by providing students with feedback concerning their explanations of sentences in text (McNamara, Levinstein, & Boonthum, 2004).

Tutorial NLD appears to be more feasible to the extent that the tutoring strategies follow what most human tutors do, as opposed to strategies that are highly sophisticated (Graesser & Person, 1994). Most human tutors anticipate particular correct answers (called *expectations*) and particular misunderstandings (*misconceptions*) when they ask the learners questions and trace the learner's reasoning. As the learner articulates the answer or solves the problem, this content is constantly being compared with the expectations and misconceptions. The tutor responds adaptively and appropriately when particular expectations or misconceptions are expressed. This tutoring mechanism is called *expectation and misconception tailored dialogue* (EMT dialogue), the mechanism incorporated in AutoTutor (Graesser et al., 2005). The EMT dialogue moves of most human tutors are not particularly sophisticated from the standpoint of ideal tutoring strategies that have been proposed in the fields of education and artificial intelligence. For example, analyses of human tutoring have revealed that tutors rarely implement intelligent pedagogical techniques such as *bona fide* Socratic tutoring strategies, modeling-scaffolding-fading, reciprocal teaching, building on prerequisites, and

diagnosis/remediation of deep misconceptions. Instead, tutors tend to coach students in constructing explanations according to the EMT dialogue patterns. Fortunately, the EMT dialogue strategy is substantially easier to implement computationally than are the sophisticated tutoring strategies.

AutoTutor (Graesser et al., 2005) attempts to hold a mixed-initiative dialogue with the student during tutoring. AutoTutor segments the student's turns into speech act units and then assigns these units into categories, such as Assertion, Short Answer, Metacognition (*I don't follow?*), Metacommunication (*What did you say?*), Definition Question (*What does X mean?*) and so on. There are approximately 20 categories of student speech acts; 16 of these are different categories of student questions. AutoTutor attempts to accommodate any student question, assertion, comment, or extraneous speech act. AutoTutor needs to produce language in addition to comprehending language. Each turn of AutoTutor requires the generation of one or more dialogue moves that either adaptively respond to what the student just expressed or that advance the conversation in a constructive fashion that answers the main question or problem. The dialogue moves within a turn are connected by dialogue markers (*Okay, Next consider*…). Some dialogue moves are very responsive to the student's preceding turn, such as the short feedback (positive, neutral, versus negative), the answers to student questions, and corrections of student misconceptions. Other dialogue moves push the dialogue forward in an attempt to cover the expectations in an answer to the main question. These forward-directed dialogue moves include Pumps (e.g., *Tell me more, What else?*), Hints, Prompts for specific words or phrases, and Assertions. The responsive and forward-directed dialogue moves together provide a mixed-initiative dialogue in which both

parties of the conversation exert an influence over the conversation. These are not scripted conversations, but rather are dynamically emerging exchanges.

AutoTutor and human tutors attempt to get the learner to fill in words and propositions in the expectations. For example, suppose an answer requires the expectation: *the force of impact will cause the car to experience a large forward acceleration.* The following family of prompts is available to encourage the student to articulate particular content words in the expectation:

1. The impact will cause the car to experience a forward _____?
2. The impact will cause the car to experience a large acceleration in what direction?
3. The impact will cause the car to experience a forward acceleration with a magnitude that is very _____?
4. The car will experience a large forward acceleration after the force of _____?
5. The car will experience a large forward acceleration from the impact's _____?
6. What experiences a large forward acceleration?

The particular prompts that are selected are those that fill in missing information if answered successfully. That is, the dialogue management component adaptively selects hints and prompts in an attempt to achieve pattern completion. The expectation is covered when enough of the ideas underlying the content words in the expectation are articulated by the student so that the expectation is sufficiently covered. LSA and other semantic analyzers determine whether the student has sufficiently articulated each particular expectation.

Evaluations of AutoTutor have been encouraging in several respects. First, AutoTutor is useful because students learn about computer literacy and physics much

better than reading a textbook for an equivalent amount of time, and nearly as well as an expert human tutor. Second, the conversations of AutoTutor are surprisingly smooth because bystanders in a bystander Turing test cannot tell whether a randomly selected turn was generated by AutoTutor or a human tutor. Third, the LSA based judgments of whether a sentence-like expectation was covered in the dialogue is approximately as accurate as a graduate research assistant. These successes are surprising because AutoTutor does not really understand the learner at a deep level, with a fine-grained alignment of knowledge states in a common ground. This raises questions about the notion of common ground. That is, do participants in a dialogue really need to know a great deal of what each other knows for successful conversation to proceed?

Closing Comments

This chapter has reviewed progress that has been made in developing computational models of text comprehension and two-party dialogue. Sufficient progress has been made in the fields of discourse processes, cognitive science, and computational linguistics to build detailed models of how discourse is comprehended and produced at multiple levels. Many of these levels are sufficiently well specified to automate them on computer. Unlike 10-20 years ago, we have reasonable solutions to handling problems of world knowledge, the vagueness and underspecification of natural language, and the management of longer threads of discourse. The computational models have evolved to the point of building useful computer technologies, such as essay graders, automated conversational tutors, question answering systems, and text analyzers that go well beyond readability formulae. It is hard to imagine what breakthroughs will emerge during the next 10 years. Some of the difficult challenges for the future, which we could not cover

in this chapter, will be computational models that perform automatic text generation, discourse-sensitive speech recognition, speech generation with appropriate intonation, and management of dialogue among three or more discourse participants.

References

Allen, J. (1995). *Natural language understanding*. Redwood City, CA: Benjamin/Cummings.

Baayen, R. H., Piepenbrock, R., & Gulikers. L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Biber, D. (1988). *Variations across speech and writing*. Cambridge, MA: Cambridge University Press.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics, 21*, 543-566.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, and discourse. *Discourse Processes, 25*, 211-257.

Charniak, E. (2000). A maximum-entropy-inspired parser. *Proceedings of the First Conference on North American Chapter of the Association for Computational Linguistics* (pp. 132-139). San Francisco, CA: Morgan Kaufmann Publishers.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T. & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471-533.

Clark, H.H. (1996). *Using language*. Cambridge: Cambridge University Press.

Core, M., Ishizaki, M., Moore, J.D., Nakatani, C., Reithinger, N., Traum, D., & Tutiya, S. (1999). The report of the third workshop of the Discourse Resource Initiative, Chba University and Kazusa Academia Hall. Technical Report No. 3, Chiba Corpus Project, Chiba, Japan.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

D'Andrade, R.G., & Wish, M. (1985). Speech act theory in quantitative research on interpersonal behavior. *Discourse Processes, 8*, 229-259.

Fellbaum, C. (1998) (Ed.) *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Fox, B. (1993). *The human tutorial dialogue project*. Hillsdale: Erlbaum.

Glenberg, A.M. (1997). What memory is for. *Behavioral and Brain Sciences, 20*, 1-19.

Graesser, A.C., & Clark, L.C. (1985). *Structures and procedures of implicit knowledge*. Norwood, NJ: Ablex.

Graesser, A.C., Gernsbacher, M.A., & Goldman, S.R. (2003). Introduction to the Handbook of Discourse Processes. In A.C. Graesser, M.A. Gernsbacher, and S.R. Goldman (Eds.), *Handbook of discourse processes* (pp. 1-24). Mahwah, NJ: Erlbaum.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36*, 193-202.

Graesser, A.C., Olney, A., Haynes, B.C., & Chipman, P. (2005). AutoTutor: A cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue. In C. Forsythe, M.L. Bernard, and T.E. Goldsmith (Eds.), *Cognitive systems: Human cognitive models in systems design.* Mahwah, NJ: Erlbaum.

Graesser, A.C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104-137.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-95.

Gratch, J., Rickel, J., Andre, E., Cassell, J., Petajan, E., & Badler, N. (2002). Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, *17*, 54-63.

Grosz, B.J., & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12* (3), 175-204.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora, *Proceedings of the Fourteenth International Conference on Computational Linguistics.* Nantes, France: ACL.

Jurafsky, D., & Martin, J.H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Klare, G.R. (1974-1975). Assessing readability. *Reading Research Quarterly, 10*, 62-102

Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Landauer, T, McNamara, D.S., Dennis, S., Kintsch, W. (in press)(Eds.), *LSA: A road to meaning.* Mahwah, NJ: Erlbaum

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal coreference resolution. *Computational Linguistics*, *20,* 535-561.

Larsson, S. & Traum, D. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering***,** *6* (3-4), 323-340.

Lehmann, F. (1992)(Ed.). *Semantic networks in artificial intelligence.* Oxford, England: Pergamon Press.

Louwerse, M.M., & Mitchell, H.H. (2003). Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes, 35*, 199-239.

McNamara, D.S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-287.

McNamara, D.S., Levinstein, I.B. & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers, 36*, 222-233.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. *Proceedings of the 18th International Conference on Computational Linguistics* (pp. 869-875). Montreal, Canada.

Moore, J. D. (1995). *Participating in explanatory dialogues*. Cambridge: MIT Press.

Pennebaker, J.W., & Francis, M.E. (1999). *Linguistic inquiry and word count (LIWC)*. Mahwah, NJ: Erlbaum.

Pickering, M.J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Brain and Behavioral Sciences, 27*, 169-190.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223-255). New York: Academic Press.

Rich, C., Sidner, C. L., & Lesh, N. (2001). COLLAGEN: Applying collaborative discourse theory to human-computer interaction. *AI Magazine*, *22*(4), 15-25.

Roy, D. (2005). Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences, 9*, 389-396.

Rus, V. (2004). A first exercise for evaluating logic form identification systems, *Proceedings Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3),* at the Association of Computational Linguistics Annual Meeting, July 2004. Barcelona, Spain: ACL.

Schank, R. and Riesbeck, C. (1981). *Inside computer understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Schegloff, E.A., & Sacks, H. (1973). Opening up closings. *Semiotica, 8*, 289-327.

Schober, M. F., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 60*, 576-602.

Sekine, S., & Grishman, R. (1995). A corpus-based probabilistic grammar with only two nonterminals. *Fourth International Workshop on Parsing Technologies (pp.* 260-270). Prague/Karlovy Vary, Czech Republic.

Sinclair, J.M., & Coulthard, R.M. (1975). *Towards an analysis of discourse: The English used by teachers and their pupils.* London: Oxford University Press.

Sowa, J.F. (1983). *Conceptual structures: Information processing in mind and machines.* Reading, MA: Addison-Wesley.

Stevenson, M. (2002). Augmenting noun taxonomies by combining lexical similarity metrics. *Proceedings of The 17th International Conference on Computational Linguistics.* Taipei, Taiwan: ACL.

Van den Broek, P., Virtue, S., Everson, M.G., & Tzeng, Y., & Sung, Y. (2002). Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. In J. Otero, J. Leon, & A.C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 131-154). Mahwah, NJ: Erlbaum.

van Dijk, T.A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

VanLehn, K., Jordan, P., Rosé, C. P., Bhembe, D., Bottner, M., Gaydos, A., et al. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S.A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring* (pp.158-167). Berlin: Springer–Verlag.

Veloso, M.M., & Carbonell, J.G. (1993). Derivational analogy in PRODIGY: Automating case acquisition, storage, and utilization. *Machine Learning, 10*, 249 – 278.

Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*, 36-45.

Winograd, T. (1972). *Understanding natural language.* New York: Academic Press.

Zwaan, R.A., & Radvansky, G.A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.