DISCOURSE AND INFERENCE

Jerry R. Hobbs

Information Sciences Institute
University of Southern California
Marina del Rey, California

September 5, 2003

Chapter 1

The Role and Structure of a Theory of Discourse

1.1 The Problems of Discourse

Edward Sapir begins his book *Language* with these two sentences:

> Speech is so familiar a feature of daily life that we rarely pause to define it. It seems as natural to man as walking, and only less so than breathing.

We read these sentences without difficulty, and we would hear them with similar ease. Yet when examined more closely, this text poses a wealth of problems.

Consider the pronoun "it" at the end of the first sentence. How do we know it refers to speech rather than daily life? We cannot tell just from the syntactic structure, for the following sentence has the same syntactic structure, and we resolve the pronoun the other way:

> Secrecy is so pervasive a feature of our foreign policy that it is hard to learn anything about it.

No simple view of semantics helps us here either. It is just as possible to define daily life as it is to define speech. The answer apparently involves knowing a complex relationship between the familiarity of something and our lack of awareness of it. But how do we access this knowledge and use it to resolve the pronoun reference?

Or consider the word "of". "Of" can mean many things – possession of a physical object, as in "a book of John's"; identity, as in "the city of Boston"; participation in an event, as in "the destruction of Rome"; and other more complex relations. What does it mean here? A feature must be a feature of something, and daily life is that something; this is what "of" conveys. But how do we know this? The word "feature" alone does not give us the answer, for one can say,

> Speech has the feature of familiarity,

in which "of" is the "of" of identity.

Sapir presupposes that the reader knows speech is a feature of daily life, but what kind of relationship has to hold between two things, such as speech and daily life, for one to be a feature of the other? The phrase "so familiar a feature of daily life that..." makes a comparison along a dimension defined by the words "familiar a feature of daily life". What is the specific nature of that dimension, and how are we able to locate a place for speech on it. If we did frequently pause to define something, what would that say about where it lay on this dimension?

This by no means exhausts the problems presented by the first sentence. The word "speech", for example, is lexically ambiguous between language and a formal address that someone delivers. The syntax of the phrase "so familiar a feature of daily life that..." is quite unusual. We might analyze it as an adjective followed by a noun phrase, but not just any noun phrase will do. We can't say, for example, "so familiar some feature of daily life that...." What is the general rule? By "we" does Sapir mean just himself and his reader? The set of all linguists? The set of all people? Does it matter whom he is referring to, and if it doesn't, how do we know it doesn't? For there are many times when it does matter. What is the implicit relation between day and life that is encoded in the phrase "daily life"? We can ask about "rarely" how many times counts as rarely and what is "pausing to define something" that it can occur rarely. What is speech that it can be defined? In what sense, if any, does pausing enable us to define speech?

The second sentence presents the same wealth of problems: How do we know "it" again refers to "speech"? What are speech, walking and breathing that they can be natural? How does the phrase "to man" alter the meaning of "natural"? What is the nature of the dimension "naturalness to man" that allows us to compare speech, walking and breathing on it? How are we able to construct this dimension for comparison, seemingly so effortlessly. What does Sapir mean by "only" – that breathing is the unique thing that

is more natural to man than speech, or does "only" mean something more like "slightly" or "just"? What predicate is the word "so" standing in for? What function does the word "seems" play, and why was it chosen rather than "is"?

It is not just the words and phrases in single sentences that present difficulties. The very fact that two sentences occur next to each other in a text, and sound right together, requires explanation. What is the relationship between Sapir's two sentences that we are not surprised that they are part of the same text? Intuitively, we might say that they elaborate on the same theme. Pressed to be more specific, we might say that both sentences are making comparative statements about the familiarity or naturalness of speech. How do we see past the words used to this similarity of expressed thought, and why does this similarity seem to confer a feel of coherence to the text?

Finally we might ask why Sapir is telling us this? Why did he write these two sentences? It is not new information. Indeed, it is clear he expects us to agree with what he says. But as soon as we agree with it, it suddenly becomes less true. His writing it and our reading it have suddenly made speech seem less familiar and less natural, and this appears to be Sapir's goal—to make speech seem suddenly problematic. The next sentence in the text is "Yet it needs but a moment's reflection to convince us that this naturalness in speech is but an illusory feeling." The title of the first chapter is "Introductory: Language Defined", and the purpose of the entire book *Language* is to subject speech to close scrutiny. In short, Sapir wrote his first two sentences for very much the same reasons that I wrote my first paragraph—to set up the first half of a contrast, the second half of which makes an investigation, and a book, possible.

I chose the Sapir text for a number of reasons—its ecological validity, its good but nonliterary style, the prestige of its author, the possibilities of self-reference—but not for the discourse problems it posed. It is not at all unusual in this regard. Virtually any two consecutive and moderately complex sentences would have served as well. With all of the difficulties presented by ordinary texts, how is it that we understand them so readily? More seriously, how is it that we understand them at all?

From one point of view, the point of view of daily life, all of these problems are not really problems. How do we solve them? We just do. But if we ask about the mechanisms that underlie our ability to solve these problems, and we are very strict about what we mean by "mechanism", all of these become significant problems indeed.

For a number of years, researchers in the fields of artificial intelligence (AI) and cognitive psychology have looked at discourse from just such a point of view. The theory developed in this book arises out of this AI tradition. Workers in AI have tried to build natural language understanding systems, and the computer has placed quite stringent limits on what counts as a mechanism. This work is of more than just technological relevance. Today, only people understand language. Efforts to get computers to understand language must therefore draw on what is known about how people do it. Conversely, in spite of the substantial differences in architecture and power between the human brain and the present-day computer, there is embedded within most natural language processing systems a theory, implicit or explicit, of how language would be comprehended by any intelligent entity, hence by humans. If a theory of discourse is framed at a sufficiently abstract level, it may apply equally well to human and computer understanding.

The fundamental lesson so far of this AI research is that we understand discourse so well because we know so much. We understood the Sapir text because we know a lot about speech, familiarity, features, daily life, and so on. But we do not just *have* the knowledge, we are able to *use* the knowledge to make sense of the text. The central problems in understanding how people interpret discourse are therefore how to characterize the knowledge that people have and the processes they use to deploy that knowledge in the task of interpreting discourse. It is the aim of this book to address these two problems.

Before going on, it will be useful to fix certain crucial parts of our vocabulary. As a start, "discourse" will be taken to mean people communicating with each other, although in the next section we will focus in more on our immediate concerns. A "text" is a fragment of discourse produced by a single speaker. A fragment of discourse in which more than one participant speaks is referred to as a "dialog". The words "speaker" and "listener" will often be taken to cover writers and readers as well. The word "listener" is used rather than the more common "hearer" because listening is a more active process than hearing. The mapping from some representation in a speaker's mind into a text will be referred to as "production" rather than "generation", since the latter has acquired a very specific technical meaning in linguistics. The mapping from the text into a representation in the listener's mind will be referred to as "interpretation" rather than "compre-

hension", since toward the end of the book I wish to divorce the process from individual cognition.

1.2 The Locus of This Inquiry in the Study of Discourse

Discourse can be viewed in many ways. A text is concocted in the mind of one person. It is produced and exists in the world as something that can be examined independently. It is interpreted, perhaps in different ways, by each member of the speaker's audience, and takes on a life in each listener's mind. The text carries meaning—it bears some quite detailed relation to the world. In fact, it carries more than one meaning: the speaker means one thing by it, each of the listeners might take it to mean something quite different, and it has a meaning independent of its occasion which is imposed by the language in which it is framed. By exchanging texts, people increase their knowledge and increase the portion of their knowledge that they share, thereby binding themselves more tightly together, increasing for example their possibilities for joint action. The texts people exchange take their place in a long line of similar texts exchanged by similar people, and this discourse is the principal means by which social organization is constituted and given continuous life.

In this inquiry we will confine ourselves primarily to one view of discourse—its interpretation. We will investigate the means by which a listener transforms the text, as something out in the world, into something in his[1] own mind. That is, we will study the psychology of discourse, rather than the philosophy, the sociology, or the history.

A speaker's production of a text is also a psychological process, but this volume will be for the most part concerned with interpretation. The two processes certainly access the same knowledge, and they surely overlap importantly in their subprocesses. For example, in production we frequently ask ourselves how we can say something in a way that our listener will be able to interpret it, and in interpretation we frequently ask ourselves what the speaker is really trying to say. But in this volume the focus will be on interpretation, with production viewed, when at all, as a process of con-

[1]Rather than bloat this book with the disjunction "he or she", I will uniformly refer to the generic speaker as "she" and the generic listener, or hearer, as "he". This should be easy to remember since "hearer" starts with "he" and "speaker" and "she" both start with "s". The occasional other generic characters in this book will be referred to as "he or she".

structing the utterance in a way that will produce the right interpretation.

Much of this study will consider a single listener interpreting a single text produced by a single speaker. But in addition we will examine some multi-participant discourses. Something very different happens when more than one speaker is involved than when the discourse is largely under the control of a single speaker. With more than one speaker, there are likely to be more than one set of intentions or goals for the discourse. These intentions will come into conflict and the conflicts will be resolved in complex and unpredictable ways. To study multi-participant discourse, we need all the machinery we need for single-speaker discourse, plus a way of talking about and analyzing the participants' conflicting plans. Nevertheless, each participant must interpret the contributions of the others, and although often at a much finer grain, the interpretation processes are the same as in interpreting single-speaker discourse.

In discourse, information is communicated from one person to another by some means. The communication can be verbal or nonverbal; it is usually a mixture of both. Much of the work in the AI tradition is based on the notions of the logical representation of knowledge and language and of the propositional content of sentences. In nonverbal communication, including intonation, it is not clear a priori to what extent these notions apply. What is the propositional content of a gesture? In this book we focus primarily on verbal communication, but nonverbal communication will be addressed in a very limited way. In any case, we can be sure that verbal communication can be meaningfully investigated in isolation. The success of written communication guarantees this: we are able to understand discourse in the absence of its nonverbal component.

Communication may also be spoken or written, and it may vary along a large number of other dimensions. These distinctions, however, will not be significant in our inquiry. Most of the specimens examined in this book are written texts, and there is perhaps a bias toward written discourse. Written texts have a number of attractive features for this inquiry. A written text is under the complete control of the author; the reader does not have a chance to intervene in the course of the writing to redirect it (although a writer may imagine her reader's responses as she writes). Except for illustrations, it is entirely verbal. The reader's interpretation takes place in isolation from the production of the text, and thus is a separable phenomenon amenable to independent investigation. But this study is decidedly not restricted to written discourse. There are certainly numerous differences between written discourse and the verbal component of spoken discourse, as research has

shown (e.g., Tannen, 1982). But the analysis of a wide variety of materials, both spoken and written, have indicated that these differences do not prove especially significant for the principal focus of this book, the processes for deploying knowledge in interpretation.

Among the other dimensions along which discourses can be classified are the spatial or temporal proximity of the participants, the degree of formality, the amount of prior rehearsal, and the extent of the participants' shared knowledge (cf. Rubin, 1978). The place of a discourse on each of these dimensions has many influences on the shape of the discourse—lexical choice, fullness of the definite descriptions, and so on. Many of the differences arise out of the differences in shared knowledge each of these conditions imposes and differences in how exacting or forgiving the listener should be. But once the extent of shared knowledge and the exactitude of the interpretation are fixed, it is likely that the procedures that use this shared knowledge to achieve an interpretation do not differ significantly from condition to condition. For this reason, these distinctions do not play a role in the present investigation.

To summarize, the phenomenon under investigation will be the processes involved in interpreting stretches of verbal communication produced by a single speaker, sometimes in dialogue with other participants. It is obvious that a theory of this phenomenon will play an important role in any more complete theory of human discourse.

This investigation is organized around seven "target texts". These are

1. the first two sentences of Edward Sapir's *Language*;
2. the first paragraph of Carson McCullers' *The Ballad of the Sad Cafe*;
3. the first two paragraphs in an article from the San Jose *Mercury News* business section;
4. two paragraphs from a *Science* magazine article on AIDS;
5. two brief military equipment failure reports;
6. Shakespeare's 64th sonnet;
7. the transcript of about a minute and a half of a three-person decision-making meeting.

These target texts are given in an appendix. In this book I hold myself responsible for the representational issues, the syntactic constructions, the underlying commonsense knowledge, and the interpretation problems presented by these texts. The aim of the diversity of the texts is to force generality in the theoretical framework. If all of the phenomena in these texts can be handled in a uniform fashion, that is a good argument for the broad applicability of the theory.

Each of these texts comes from a larger corpus that will be used at several points in the investigation as "target corpora". They are

1. the first two paragraphs of Edward Sapir's *Language*;
2. the first 14,000 words of Carson McCullers' *The Ballad of the Sad Cafe*;
3. about 30,000 words from the San Jose *Mercury News* business section;
4. four *Science* magazine articles on AIDS (nearly 22,000 words);
5. thirty-six brief military equipment failure reports;
6. all 154 of Shakespeare's sonnets (nearly 18,000 words);
7. the transcripts of about 68 minutes of five three-person decision-making meetings (nearly 14,000 words).

In addition, the lyrics of the one hundred most popular country-and-western songs of the 1980s are used as a target corpus. Again this diversity forces generality.

In addition, I will make use of various naturally occurring fragments of discourse, as well as made-up linguistic examples, where they illustrate the problem being discussed. But they will not play the same forcing role as the target texts and the target corpora.

1.3 Methodological Aims and Assumptions

In this section I attempt to present a brief but coherent theoretical and methodological statement, but in doing so I find it necessary to race headlong through some hotly contested territory, imagining attacks from every theoretical orientation as I go. There is hardly a reader who will not find objectionable some position that seems to be taken without further justification in a casual phrase. I can only plead that I have tried to choose

my phrases carefully and that the positions I adopt, though controversial, have been adopted and argued at length by others and could be defended.[2] Since one's theoretical stance tends to permeate one's work in subtle ways, I feel it is necessary to lay out the presuppositions of this enterprise at the beginning, even though an adequate treatment of these issues would require a volume in itself.

A scientific theory is a more or less formal explanation of more or less extensive data. Formality promotes intellectual honesty in what the theory predicts. Confirmation and falsification result from the comparison of the theoretical prediction with the data and tell us what data is and is not covered.[3] A theory is ultimately judged by the elegance of its formal explanations and the coverage of its data. Idealization is a way of trading in extent of coverage for elegance in explanation.

This paragraph merits repetition.

A scientific theory is a more or less formal explanation of more or less extensive data.[4] Formality promotes intellectual honesty in what the theory predicts. Confirmation and falsification result from the comparison of the theoretical prediction with the 'data' and tell us what 'data' is and is not covered. A theory is ultimately judged by the elegance of its formal explanations and the coverage of its 'data'. Idealization is a way of trading in extent of coverage for elegance in explanation.

In presenting a theory of discourse, it is necessary to state at the outset what is to count as an explanation, what makes an explanation formal, and what data one is willing to be held responsible for. Each of these is taken up in turn.

Explanation in AI and cognitive psychology is based on the computer metaphor: whatever else it may be, the brain is at least a kind of computer. We seek to understand how psychological processes could be implemented in terms of the symbol manipulation operations of computability theory. Insofar as we succeed, we will say we have "explained" them. There are several reasons this account of explanation is compelling.

The first is simply that it is often useful to try to understand one complex

[2] Perhaps the best statement of the theoretical stance of cognitive science is that of Haugland (1981). Other good accounts can be found in Newell and Simon (1976), Dennett (1978), Pylyshyn (1981), and Winograd (1977).

[3] "Data" is a singular English mass noun derived from a plural Latin count noun.

[4] Or what Lakatos (1970) might call "'data'", conventionally agreed upon.

system by comparing it with another. Analogies elucidate. This is particularly true when different aspects of the two systems are open to inspection. Large computer programs are among the most complex objects which (in principle) are entirely under our control, and they exhibit behavior that at least superficially resembles some aspects of intelligent human behavior. The analogy between cognition and computation is fruitful for this reason alone.

A second reason for adopting the computer metaphor is purely technological. We know quite well how the level of symbol manipulation can be implemented in electronics, and thus whatever success we have can at least lead to useful computer systems. As Terry Winograd has said, if in trying to build an airplane, we end up building a boat instead, we'll go for a sail.

But the primary motivation for the computer metaphor is the promise it holds out for "reducibility".

Science is organized by levels, a strategy that is successful probably because nature is organized by levels. There are several ways we can view these levels. First, we can view them as *levels of description.* Nature cannot usefully be described solely in terms of the motions of elementary particles. We have found it convenient to define or hypothesize larger-scale entities and to couch our theories in terms of them. We then try to account for the behavior of these entities in terms of the entities provided by the theory of the phenomena one or two levels down. Thus, chemists seek to understand in quantum theoretic terms why molecules react as they do.

We can also view the levels as *levels of organization.* That is, they are not merely convenient fictions that allow our poor, finite minds to understand what is going on. There is something in nature that actually corresponds to these large-scale entities and actually behaves approximately in the manner that our theories describe. The argument for assuming these things are really out there in the world is what has often been said: We should adopt the ontology implied by our most successful theories. The reality of the ontology is the best explanation for the success of the theory. Molecules, cells, tissues and organs, organisms, herds, and nations are not merely stories we tell. They really do exist.

Evolution has proceeded by levels of organization because these represent *levels of competence.* Each level is characterized by the achievement of stable forms, out of which larger structures can eventually be constructed. Molecules are stable forms constructed out of atoms; stars, rocks and cells are stable forms constructed out of molecules; multicellular animals, including people, are stable forms constructed out of cells; and social organizations

are stable forms constructed out of people.

There is much controversy about the ways in which a scientific theory of one level is or ought to be "reducible" to that of another level. Most arguments against the reducibility of a higher-level science, or "macro-science", to a lower-level science, or "micro-science", take the following form: the macro-science and micro-science each require idealizations for their most elegant formulations, and there may simply be a mismatch between their idealizations. For example, physiology concerns itself primarily with the prototypical members of a species, whereas population biology must concern itself with deviations from the prototype (Dupré, 1983). It is not just that elegance in the macro-science would be sacrificed. Usually, computational complexity precludes the reduction (although the reduction of the thermodynamics of an ideal gas to statistical mechanics is a notable exception, where computational complexity does not preclude the reduction). Moreover, the structure and behavior of an entity at the higher level cannot be explained by describing *only* that entity at the lower level and not the environment with which the entity interacts.[5] Thus, one cannot describe the structure and life history of a rock on the basis of its mineralogy alone without reference to tectonic influences, and similarly, psychological phenomena depend on sociological phenomena as well as physiological principles. Finally, it is often theoretically impossible even to *state* important global properties of large systems in the terms provided by scientific theories of their components (Davidson, 1981; Moore, 1980). Such arguments show that laws of the macro-science cannot be replaced by complex statements in the language of the micro-science and proven as theorems from lower-level axioms. The macro-science generally concerns itself with emergent entities whose boundaries become very fuzzy when unpacked into the entities of the micro-science. Prediction does not become possible in the macro-science, resident on the laws of the micro-science. The entities of the higher level are generally very complex dynamic systems of entities at the lower level, and although gross regularities may be established, the fine details of higher entities and processes cannot be derived. We understand the underlying physics of rivers, hurricanes, and volcanoes, but we can't predict their behavior, except within very coarse limits.

But this would be a very strong form of reducibility, one that might be called "replacibility".

It is nevertheless obviously true that, say, geological processes are "im-

[5]This is a straw man Putnam (1981) does battle with.

plemented" or "realized" as complex chemical and physical processes, and how they are implemented is an important question in its own right. An answer to that question constitutes another kind of reducibility, a "reducibility in principle". This is stronger than Davidson's (1981) notion of "supervenience", which means only that there is no change in properties at the higher level without a corresponding change at the lower level. In this sense, intentional psychology is surely supervenient upon neurophysiology. But the situation in geology is quite different. We know, in principle, how the behavior of a geological entity, such as a river, depends on chemical properties, such as the viscosity of water and of various minerals, and on the laws of physics. We really do have a story to tell.

There is a stage that some sciences have gone through and others have not, that represents a qualitative advance. It happens when it is understood, at least (and generally no more than) in principle, how the entities and processes at one level emerge from entities and processes at lower levels. Geology passed through this stage in the 1960s with the widespread acceptance of plate tectonics. Before then, ever since the eighteenth century, explanation in geology bottomed out in a mysterious process of "uplift". Plate tectonics explained uplift in terms of underlying physical processes. At that point, geology became, in the sense of reduction I intend here, a "reduced" science.

In this sense, it is desirable for psychology to provide a reduction in principle of intelligent, or intentional, behavior to neurophysiology. Because of the extreme complexity of the human brain, more than the sketchiest account is not likely to be possible in the near future. Nevertheless, the central metaphor of cognitive science, "The brain is a computer", gives us hope. Prior to the computer metaphor, we had no idea of what could possibly be the bridge between beliefs and ion transport. Now we have an idea. In the long history of inquiry into the nature of mind, the computer metaphor gives us, for the first time, the promise of linking the entities and processes of intentional psychology to the underlying biological processes of neurons, and hence to physical processes. We could say that the computer metaphor is the first, best hope of materialism.

The jump between neurophysiology and intentional psychology is a huge one. We are more likely to succeed in linking the two if we can identify some intermediate levels. A view that is popular these days identifies two intermediate levels—the symbolic and the connectionist.

Intentional Level
|
Symbolic Level
|
Connectionist Level
|
Neurophysiological Level

The intentional level is implemented in the symbolic level, which is implemented in the connectionist level, which is implemented in the neurophysiological level.[6] The aim of cognitive science is to show how entities and processes at each level emerge from the entities and processes of the level below.

The reasons for this strategy are clear. We can observe intelligent activity and we can observe the firing of neurons, but there is no obvious way of linking these two together. So we decompose the problem into three smaller problems. We can formulate theories at the symbolic level that can, at least in a small way so far, explain some aspects of intelligent behavior; here we work from intelligent activity down. We can formulate theories at the connectionist level in terms of elements that behave very much like what we know of the neuron's behavior; here we work from the neuron up. Finally, efforts are being made to implement the key elements of symbolic processing in connectionist architecture. If each of these three efforts were to succeed, we would have the whole picture.

In my view, this picture looks very promising indeed. Mainstream AI and cognitive science have taken it to be their task to show how intentional phenomena can be implemented by symbolic processes. The elements in a connectionist network are modeled closely on certain properties of neurons. The principal problems in linking the symbolic and connectionist levels are representing predicate-argument relations in connectionist networks, implementing variable-binding or universal instantiation in connectionist networks, and defining the right notion of "defeasibility" in logic to reflect the "soft corners" that make connectionist models so attractive. Progress is being made on all these problems.

Although we do not know how each of these levels is implemented in the level below, nor indeed *whether* it is, we know that it *could* be, and that at least is something.

[6] Variations on this view dispense with the symbolic or with the connectionist level.

At the present time our computational models of mental processes fall far short of what people are capable of. But this in no way constitutes an argument against the computer metaphor. Whatever the limits of the computer metaphor are, we have not begun to approach them yet.

Should the enterprise succeed, there is no reason to feel that it would diminish in any way our image of humankind, any more than the view of the human body as a chemical and mechanical device has devalued the body. A theory that did turn out to be dehumanizing would simply be wrong; it would contradict the facts as we know them. So far, however, the evidence has all gone in the other direction, showing us, if the brain is a computer, what a magnificent computer it must be. The computer metaphor does not pose a threat to humanistic psychology, for as Boden (1977) has pointed out, it *is* a variety of humanistic psychology. It allows us to view people scientifically not merely as exhibiting behavior but as taking action.

When we adopt the computer metaphor, we are adopting a particular, very stringent requirement for explanation in our theory. An explanation is a specification of computable procedures that will produce the behavior under investigation. If what we provide isn't computable, it isn't an explanation, only a prose description, and most of the point of appealing to the computer metaphor is lost. This requirement is what Wilensky has called "procedural adequacy" (Schank and Wilensky, 1977).

This requirement places those who study discourse on the horns of dilemma. On the one horn, we would like our theories to be faithful, empirically adequate accounts of the way people actually process discourse, the knowledge they actually use, and the goals they are actually driven by; on the other, we require computable procedures that actually solve the discourse problems we are faced with. In most cases, we can't have both. If we adhere to empirical adequacy and do not go beyond what our data warrants, we will not solve the problems, for too much is going on that is simply unobservable. We will be condemned to sterile theories. When we try to construct procedures that work, we are on shaky empirical grounds. Computability forces us to specify procedures to a level of detail beyond what is justified by solid evidence. It is for this reason that work in AI often seems *ad hoc* from a cognitive point of view.

There is at least a partial escape from *ad hoc* theories, that workers in AI have not always availed themselves of. This is to frame the explanations at

as abstract a level as possible, while still retaining computability. Workers in AI sometimes[7] couch their theories in terms very close to the actual code of some implemented system. But, in spite of the power of the computer metaphor, the brain and the present-day computer are sufficiently different that a discussion at this level of detail is of no psychological interest, whatever its technological merits. In line with current practice in AI, the theory presented here is expressed in formal logic. It thus retains computability while being abstracted away from implementational details, such as data structures and serial versus parallel computer architectures. The most any AI, or other "top down" (cf. Dennett, 1978), approach in psychology can hope to do is to discover a *possible* explanation of behavior. By making the formalism maximally noncommittal on inessentials by using formal logic, we expand the range of possibilities the theory marks out. Rather than running counter to psychological reality, as often assumed, formal logic enhances the psychological relevance of one's theories.

Now the data—what is to be our concern and what is not. Psychological data is abundant. We are flooded with masses of it everyday as we interact with people, hear them talk, and observe their behavior. This psychological data is available to everyone, virtually without effort. Slightly less accessible are people's reports on this behavior—for language, such things as judgments about the grammaticality of a sentence or the appropriateness of a response, and clarifications, paraphrases and expansions of ambiguous or elliptical utterances. Finally, there are various sorts of more or less exotic, hard-to-obtain data, such as data on reaction times and eye movements, and results of fMRI studies. Ultimately a psychological theory will have to be responsible for all of this data, but today it is necessary to choose a portion of the data that seems significant and coherent and looks as if it will yield a reasonable theory. This inquiry will focus primarily on the middle of the three categories of data – specifically, on *interpretation reports*, people's reports on how they have interpreted a word, a phrase, or a larger stretch of text. We cannot use the most accessible class of data—utterances—for a theory of discourse interpretation, because it is not data *about* interpretation. It is about production. There are also problems with certain of the more exotic classes of data, as discussed below.

The use of interpretation reports is the genius—and the most significant

[7] Formerly more than today.

contribution—of Chomsky's competence-performance distinction. It is not possible to build a science of *utterances*, because it is not possible to predict utterances. The mystery of choice intervenes. But it *is* possible to predict interpretation reports of a very limited sort, such as grammaticality judgments. Whereas a science of utterances would be a science of performance, a science of interpretation reports is a science of competence. It is a fundamental assumption of modern linguistics that the knowledge of language that we make use of in producing and interpreting utterances is the same knowledge of language that we make use of in interpretation reports. This assumption, that there is such a thing as linguistic competence, is what makes prediction, and hence a science of language, possible.

Our primary data will thus be certain kinds of interpretation reports. The first are reports of what a referential expression refers to, such as that "he" in the text

John can open Bill's safe. He knows the combination.

refers to John and that "the index" in

(1.1) John picked up a book and turned to the index,

refers to the index of the book John picked up. Second are brief expansions or paraphrases of such constructions as compound nominals and similar predications conveyed in sentences. Examples are reports that "wine glass" means a glass whose function is to contain wine, and that "turned to the index" in (1.1) means turning the pages of the book until finding the index, rather than, say, turning one's body to face the index.[8] Of course such reports are themselves only texts, but we will take them to be reliable and priveleged. A satisfactory theory cannot claim that a listener resolved an occurrence of "he" to Bill when the listener reports that he resolved it to John.[9]

We will also consider as psychological data to be explained the very fact that people can interpret a text and respond in an appropriate way. For example, suppose I ask someone if he could hand me a pencil, and he hands me the pencil, rather than, say, replying "Yes" or staring at his hand in

[8] A more elaborate treatment of some of these interpretation reports and others can be found in Mann et al (1975).

[9] Note however that we do not assume such reports are reliable across informants.

confusion. This is psychological data, and we may ask how such a thing could occur.

Precisely what phenomena, and consequently what sort of interpretation reports, I will attempt to account for emerges as the book progresses. But it is perhaps useful for me to admit a bias at this point. Psycholinguists are frequently concerned with discovering and modelling people's shortcomings. Among linguists this concern takes a more theoretical turn; so that their postulated mechanisms will not have capabilities people lack, they strive to constrain the power of their theories. The typical result among linguists and psycholinguists has been to show that a particular process is too powerful and fails to explain what people don't or can't do. By contrast, the typical result in AI shows that a proposed process is not powerful enough and thus doesn't explain what people can do. It is a frequent charge among linguists that the hypothesis of a particular mechanism is vacuous because it has the power of a Turing machine. But it seems completely obvious to me that a mechanism capable of understanding natural language discourse *must* have the power of a Turing machine. The question is what that Turing machine is. In fact, it is not even clear that a Turing machine is adequate; the computer metaphor really might not turn out to be appropriate. It is precisely this hypothesis that AI and cognitive psychology have taken it upon themselves to test. It is certainly the ultimate aim of any theory to account both for people's abilities and their inabilities, but it seems to me that people's ability to engage in discourse so vastly outstrips our ability to model it that the primary problem is not to constrain our theories, but to enhance their power.[10]

Interpretation reports come in two varieties—positive and negative. Some interpretations do not occur. For example, in (1.1) the listener will not normally take "the index" to refer to the index of the first book mentioned in the bibliography of the book John picked up. An adequate theory must explain both kinds of judgments. But when it comes to the interpretation of discourse, it might seem that negative judgments are hard to come by. Interpretation depends heavily on the context of utterance and the knowledge that is shared by the speaker and listener. This leaves us a lot of room to maneuver. If we exclude codes, it is unlikely that "the index" in (1.1) will

[10]I think most AI practitioners share my bias. From the point of view of AI as technology, there is something perverse about seeking to imitate people's shortcomings, except where, as Black (1980) has urged, we view people's shortcomings as indicative of greater strengths. For example, a lack of facility with center embeddings may be a side-effect of the ability to parse in real time, and thus a clue as to how the latter is done.

be taken to refer to John's pet cat, but perhaps if we were clever enough, we really could load the context in a way that would support such outlandish interpretations.

Suppose we include among our interpretation reports judgments about the acceptability of question-answer pairs. What would we say about the following pair?

A: Was he an opera buff?
B: No, he was in the war.

The coherence of these utterances is likely to elude us. This was actually overheard, however, in the following context. A was whistling an aria from an Italian opera. B remarked that his father often used to sing that, and then the above pair. So the interpretation of B's reply is that B's father learned the aria while he was stationed in Italy during World War II. People, especially linguists, are very good at imagining contexts that will turn the most bizarre collection of utterances into coherent discourse.

This might seem to pose a dilemma. Either our theory predicts a single interpretation, in which case it will be wrong in many contexts, or it predicts a (rather large) family of interpretations, in which case it is very nearly vacuous.

In fact there is no such dilemma. The interpretation procedure – call it F—is a function not of the text alone but also of some representation of the context. Texts are interpreted with respect to a knowledge base, and much of what we think of as contextual differences can be characterized as differences among knowledge bases. The knowledge we bring to bear on texts includes knowledge of the local environment and the situation of utterance, knowledge of the surrounding discourse, knowledge of what the speaker's and one's own aims are, knowledge of what knowledge one shares with the speaker, as well as general world knowledge. The order of access of this knowledge may vary according to salience and other attentional factors. Each of these aspects of knowledge will be examined in the chapters that follow. For now let us encapsulate it all in the symbol K. Then we can summarize interpretation in the following "formula":

$$F(T, K) = I \tag{1.2}$$

An interpretation procedure F is applied to a text T and a knowledge base K to produce an interpretation I.

This formula says that one cannot talk about the interpretation of a text without specifying the knowledge base that the text is interpreted with respect to. To put it another way, the context is one of the parameters of interpretation. If the context is changed, K is changed, and we would expect F to produce a different result. This is not a way of squirming out of the requirement of falsifiability. We have not thus made our interpretation process "tailorable", in the sense of van Lehn et al. (1983), to virtually any answer we desire. Once the new K is specified precisely, the requirements on the theory are still as stringent: the anomalous interpretation must be produced. Texts really are interpreted differently in different contexts, and this is a fact that requires explanation in discourse theory.

The situation is quite analogous to one Lakatos (1970) imagines in pre-relativistic celestial mechanics. An astonomer uses Newtonian mechanics and some observed initial conditions of a planet to predict its orbit. If the prediction fails to be confirmed, it does not lead him to reject Newtonian mechanics. He is more likely to postulate an unknown planet near the known one which perturbs its orbit. Similarly, if our interpretation procedure fails to predict an interpretation report, our first guess is likely to be that we used the wrong K in formula (1.2). We postulate something different in the context. This is a legitimate move, but it must be tested severely. The astronomer must postulate a specific orbit for the unknown planet, show that it explains the anomaly, and seek independent confirmation of its existence—usually with a telescope. Similarly, the discourse analyst must specify the suspected context precisely, show that F then predicts the interpretation report accurately, and seek independent confirmation of the crucial aspects of that context.

This last requirement brings up a good question however—how do we validate a hypothesized knowledge base? We will assume we have fixed upon a particular set of similar speakers, listeners, and global contexts, for which there is available a reliable source of interpretation reports. We have probably done this by picking an extensive *corpus* of texts produced by a single speaker or a coherent set of speakers, for which we may consider ourselves a part of the intended audience. A knowledge base, both its occasional and more permanent parts, and an interpretation procedure constitute a theory of the corpus, just as, in the large, physics and geology are theories of their own phenomena. The theory will generate interpretations that are formal objects, necessarily specified more precisely than can be verified by observables, but anchored in observable interpretation reports at various crucial points.

A theory of the corpus develops in a way analogous to the way in which, according to Lakatos (1970), a scientific research program develops. Lakatos distinguishes between the *basic core* of a theory—those principles that are exempt from counter-evidence—and the *protective belt* of auxilliary hypotheses that are adjusted to fit the evidence. In fact however, in a theory of a corpus, just as in a theory of natural phenomena, there is not just a two-fold division, but a set of layers. In discourse theory, when prediction fails, we are likely first to adjust our assumptions about the immediate context, then those about general knowledge. Only with greater reluctance do we modify the interpretation procedures. Nearly impervious to challenge are the overall theoretical framework and the computer metaphor itself. In general, we want to make those changes that are least consequential for the rest of our theory.[11] A sequence of such modifications yields a series of theories. A series of theories is *progressive*, in Lakatos' terms, if the successive theories cover more and more of the available evidence. In discouse, this would mean we were able to modify K, and perhaps F, in a way that correctly predicts more and more interpretation reports in more and more texts of the corpus. In validating a knowledge base and an interpretation procedure, a progressive series of theories of a corpus is the most we can hope for.

All of this makes discourse theory a difficult enterprise, but there was never any reason to suppose it would be an easy one. Nevertheless, as I hope to show in this book, a quite manageable and well-defined program of research is indicated. I will examine the nature of discourse theory in more detail in the final chapter of this book. In the meantime, the reader can view the book as laying the groundwork for such a theory of discourse, by defining precisely what F, K, I, and indeed T must look like.

Our primary data is therefore interpretation reports. But there are several commonly employed kinds of interpretation reports we will not use here. One is the reports that a subject gives of what he remembers of a text in a recall experiment. This is extremely complex behavior involving interpretation, memory, and creativity; there are many confounding factors; and there is no particular reason to believe such reports give us very direct views of memory structure, the interpretation processes, or anything else. Beaugrande (1980) gives an excellent critique of such a recall experiment.

Secondly, although it is assumed people can reliably report on their in-

[11] A further move that is available to us somewhere in this sequence is to challenge, or even dismiss, the interpretation report, thus assaulting what Lakatos calls the "interpretive theory", rather than the explanatory theory. But this is an action directed not at one's theory of the corpus, but at one's colleagues, who decide what the 'data' is.

terpretations, it is decidedly not assumed they can reliably report on their interpretation processes. Introspection does not provide data. Introspection is a mode of accessing one's intuitions, and the role of intuition in discourse theory is just what it is in any scientific or critical enterprise: it is a source of hypotheses. It has nothing to do with the validity of what is hypothesized. We process language and we have an intuitive folk theory about how we process language, and there is no reason to believe they have very much in common. There are certainly occasions when we are quite conscious of part of the interpretation process; we may mull over the meanings of certain words at length. Much else seems to happen just at the edge of consciousness. Most of what occurs, however, is probably deeply unconscious, in the sense that it is in principle inaccessible to introspection. Belief and inference play a central role in discourse interpretation theory but only as theoretical constructs, not as data to be explained, not as something that can be introspected about and reported on.[12] Correspondingly, the intuitive plausibility of a hypothesis may cause a researcher to pursue the hypothesis with special vigor, and it may win other researchers over to an approach, but it has nothing to do with the validity of the hypothesis and does not in any way constitute an argument for or against a position.

More generally, a theory based on the computer metaphor is necessarily a "deep" theory, in the sense of Moravcsik (1980); the formal explanatory machinery is typically much greater in scope than the observable data it is intended to account for. The formal machinery must produce results that can be interpreted as corresponding to the observed data to be explained, but we cannot in general expect experimental confirmation of the details of the formal process that led to these results. For example, for sentence (1.1) it is proposed in Chapter 6 that we are able to resolve the definite noun phrase "the index" by drawing an inference based on an axiom in our knowledge base that says that many books have indexes. Nevertheless, we should not expect to find any direct evidence of this inference being drawn other than a reader's report of what he resolved "the index" to.

All of this does not mean, by the way, that no processing conclusions can be drawn. Consider, for example,

[12]The confusion between inference as part of a conscious problem-solving effort and inference as a theoretical construct is common enough. It infects, for example, Zajonc's (1980) criticism of cognitive explanations in social psychology, and it even occurs, mystifyingly, in the latter parts of the otherwise excellent paper by Haugeland (1981).

(1.3) John can open Bill's safe. He ...

Who does "he" refer to? Most people will reply "John". But the complete text could have been

(1.4) John can open Bill's safe. He is going to have to get the combination changed.

Hearing this text we sense that first we resolve "he" to John, but then as the second sentence proceeds, we change the resolution to Bill, and our report about (1.3) substantiates this. This seems to indicate that we have at least two strategies for resolving pronoun references—one independent of the semantic content and the other not. This is an example of drawing rather strong conclusions about processing from people's reports about interpretations, which are fairly accessible.

Finally, there is the exotic data. It is common in cognitive psychology to seek correspondences between reaction times and complexity properties of algorithms suggested by the formal theory. But I have ignored all data on timing of mental processes. Questions of timing and complexity are very much dependent on the architecture of the machine. We know very little about the architecture of the brain, but we do know that it is parallel. Present-day computability theory has not developed sufficiently advanced formalisms for parallel computation for us to attempt to build formal models of language processing that will explain these timing results. This is not an argument against conducting such experiments and using the computer metaphor in an informal way to draw the most general possible conclusions about processing from them. But where we take the requirement of computability seriously, I think we have no choice but to set these experimental results aside for future consideration.[13]

It was said at the outset that this enterprise is a psychological one. We are now in a position to elaborate on this statement somewhat. The data that it seeks to account for—interpretation reports—is certainly psychological data. But how deep in the theory do the claims of psychological reality

[13]Pylyshyn (1980) has written an excellent analysis of the role of reaction time data in cognitive science, with which this paragraph is not, I believe, inconsistent.

go? Am I claiming, for example, that anything like formal logical expressions actually exist in the brain?

Here I am fundamentally in agreement with Chomsky's position on "psychological reality" as expressed in *Rules and Representations*, pp. 106-112. He argues that regardless of the psychological phenomena we are seeking to explain, "we should be willing to say at every stage that we are presenting psychological hypotheses and presenting conditions that the 'inner mechanisms' are alleged to meet." But I would perhaps emphasize "hypotheses" over "psychological". I think it is necessary to hold all psychological (and other) theories at arm's length. The procedures and data structures of a psychological theory are usually referred to as "cognitive processes" and "mental representations", and the justification for this is usually that one should make the ontological assumptions that one's best theory indicates. This is not an issue of fact, however, but an issue of how we will talk about things. One can certainly never expect to find direct experimental evidence for all of one's theoretical constructs, in psychology or any other science. It will generally be more convenient to speak of representations of discourse and knowledge as being in the mind and to refer to the processes of interpretation as cognitive, without embedding such talk in hypotheticals and scare quotes, although I hope not to carry this to too detailed a level. But the reader should not lose sight of the fact that when this language is used, we are in the hypothetical world of explanation and not in the "real" world of data. Whether or not a theory is psychological does not depend on its ontological assumptions.

The phrase "psychological reality" is often used as a slogan for making one's theory responsible to particular kinds of exotic data. This frequently takes the form of a call for theories of "performance" rather than theories of "competence", in Chomsky's terms. Which is intended here? There are three aspects to the notion of "competence". As discussed above, the first is the assumption that grammaticality judgments and other interpretation reports are reliable, priveleged data. This is assumed here. It is this that makes prediction possible. The second is that such reports have something to do with the listener's interpretation procedures. This is also assumed here. Otherwise, a theory of interpretation reports would not be a theory of interpretation. The third, and least important, aspect is the idealization of the "speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly..." (Chomsky, 1965). In this final sense, what is presented here is not a competence theory. Rather the interpretation procedures are explicitly made parametric on the individual listener's

perhaps idiosyncratic knowledge base. It is a performance theory in that such conditions "as memory limitations, distractions, shifts of attention and interest" have to be built into the theory somehow in order to explain how different interpretations can result on different occasions. Moreover, in the case of syntax, in Chapter 4, what is essentially a competence grammar is developed, but it is then shown how its rules are deployed in time and can yield interpretations for sometimes quite incompetent productions.

The theory aims for psychological reality in that it seeks to explain some psychological data. It falls short of psychological reality in that it does not attempt to explain *all* psychological data. But in this is does not differ from other theories.

1.4 The Structure of the Inquiry and the Book

The interpretation of discourse is a very big problem. If it is to be accessible to inquiry, we must break it into smaller subproblems, "carve nature at its joints," as the saying goes. But the saying provides a good illustration of a danger. Medicine, in carving nature at its joints, does not carve the body at its joints, with specialists in the lower right leg and the left index finger. It carves the body into coherent systems. The best way to carve the subject matter of a discipline is rarely obvious. In discourse theory one could, for example, concentrate on such "subproblems" as syntactic ambiguity, pronouns, or compound nominals, or one could confine one's inquiry to such domains as classroom discourse, telephone conversations, or children's stories. But these approaches would be the equivalent of concentrating on the left index finger. All the problems of discourse arise in each of these "simplifications", and a large-scale research program organized along such lines would result in massive duplications.

This book, and the framework informing it, is organized along quite different lines. Our goal is to present a theory of how world knowledge is brought to bear on the interpretation of discourse. Therefore, we first need a logical notation for expressing this knowledge.[14] Chapter 2 presents the logical notation that will be used in this book. There are of course numerous, quite serious problems in representing natural language concepts in logic, including time, modality, adverbials, belief and other intensional contexts, and quantification and plurality. I do not pretend to have solved all of these

[14]This is usually referred to in AI as the "representation of knowledge". The downgrading implicit in my use of "logical notation" is deliberate.

problems. Rather, my goal has been to bypass them. This is done by means of an approach that might be called "ontological promiscuity". One assumes that anything that can be talked about exists in some Platonic universe. One consequence of this is that model theory, the model theoretic semantics of the logical notation, will do virtually no work for us. All of the particulars of the meanings of various concepts will have to be encoded explicitly in the knowledge base, essentially pushing the problems from Chapter 2 to Chapter 5. On the other hand, having a uniform and relatively simple representation for all content encoded in natural language greatly simplifies the task of describing the procedures that manipulate these representations to arrive at interpretations.

Chapter 3 describes the "Interpretation as Abduction" framework that informs the rest of the book. The fundamental idea is that intelligent agents interpret their environment by finding the best explanation for the observables in it. Correspondingly, they interpret texts by finding the best explanation for the explicit content, the "observables", of the text. This picture is then subsumed under a picture in which intelligent agents interpret their environment, and texts, by finding the best explanation for why the environment, or the text, is coherent. A method of "weighted abduction" is described that has the right properties for finding a *best* explanation. It is indicated how this framework subsumes a broad range of problems in discourse interpretation; most of the remainder of the book expands on this. It is then shown how weighted abduction can be realized in a structured neural net, thereby bringing us closer to linking up intelligent behavior with neurophysiology. Finally there is a discussion of an incremental account of learning new axioms, and how this could be realized in the structured neural net model.

If to interpret a sentence is to prove abductively its propositional content, there must be a way of mapping between the string of words and the logical representation, in the notation provided in Chapter 2, of the propositional content. Syntax provides this mapping. Chapter 4 treats syntax as an elaboration of the ways in which adjacency of segments of discourse can be interpreted as predicate-argument relations. The aim in Chapter 4 is to cover all the syntactic phenomena that occur in the target texts. The result is a rather substantial subset of English grammar, modelled loosely on Pollard and Sag's Head-Driven Phrase Structure Grammar. Moreover because the target texts exhibit such "performance" phenomena as ungrammaticalities, scrambling, disfluencies, and co-constructions, these are dealt with as well. An account of is given of how a competence grammar can be deployed in

performance. The chapter closes with some remarks on modularity, and a plausible, incremental account of how syntax could have evolved.

People understand language so well because they know so much. Thus, the two major tasks in developing a theory of discourse interpretation are to specify the procedures that use knowledge to interpret discourse and to encode a sizable chunk of that knowledge. Much of the rest of the book is about the first of these tasks. Chapter 5 concerns the second. In Chapter 5 an effort is made to specify all the knowledge that is required in the understanding of all the target texts. However, the aim has been to do this by means of a systematic methodology that shows the way for extending such a knowledge base well beyond what is presented here. The knowledge is encoded at as general a level as possible, and in fact the diversity of the target texts is intended to force just this. The chapter taps into but elaborates significantly on a long tradition in lexical semantics. The most important domains that are axiomatized are abstract domains that underlie virtually every text. These include granularity, systems and the figure-ground relation, scales, change of state, space and time, causality, mental models, and goal-directed behavior. The more specific, concrete domains are built on top of these general, abstract domains, and their role in this book is primarily illustrative of how specific domains should be axiomatized.

The amount of world knowledge required for interpreting discourse is of course immense. Some take this as an argument that discourse theory is a hopeless enterprise. I do not agree. A basic premise of this work is that one can separate the knowledge that is represented and the processes that use that knowledge, and study each in isolation. To investigate the latter we do not need to know all the knowledge that is to be represented; we only need to see a representative, moderately large sampling, so that we have good intuitions as to how the knowledge would be represented and we begin so see the problems that arise in scaling up. Nor am I hopeless about the prospects of encoding vast amounts of world knowledge. Unabridged dictionaries and large encyclopedias get written, and the task of building the required knowledge base is probably not larger in scale than these efforts. My belief is that the first 10,000 axioms have to be developed with care. They will provide enough models that the next 100,000 axioms can be developed much more easily. We will then be in a position perhaps to learn the next 1,000,000 axioms automatically. My aim in this chapter is to make a serious start on that first 10,000 axioms.

Every text presents a rich set of discourse problems, as we saw in Section 1.1. The argument of Chapter 3 is that the solutions to these problems

simply fall out of the process of finding the explanation for the text, both its occurrence and its content. Chapters 6, 7, and 8 are explorations of this thesis. A broad range of discourse problems is examined in detail, and it is shown how solutions to many of the difficulties they raise are simply subsumed under the general process of abduction, or how they place constraints on its operation.

What are the discourse problems that can be posed by a text? We may divide this question into three parts—the problems that arise within single sentences (whether or not they can be solved within this narrow perspective), problems that arise when a sentence is embedded within a larger discourse, and problems that arise when the discourse has to be related to the surrounding environment. Chapter 6 examines the first class of problems. The basic building blocks of sentences are predications, by which is meant a predicate applied to one or more arguments. This implies three problems. First, what does the argument refer to? This is the coreference problem, and other problems, such as many syntactic ambiguities, can be seen as variants of it. Second, what is the predicate that is being conveyed? This includes resolving word sense ambiguities. But more generally, texts give us rather general explicit information, and we need to determine more specific interpretations. Extreme cases of this can be seen in phenomena such as compound nominals and denominal verbs, where the predicate is only implicit and must be "vivified". Third, in what way are the predicate and its arguments congruent? At the simplest level, this amounts to the checking of selectional constraints. But the issue also arises of how to interpret the predication when there is no apparent congruence. There are two deformations one can make to force an interpretation. One can assume that the argument refers not to the explicit referent but to something functionally related to the explicit referent. This is *metonymy*, and the process of interpreting metonymy is known as "coercion". Or one can assume that the predicate does not quite mean what it literally means, that is, that certain inferences normally associated with the predicate cannot be drawn in this instance. An important example of this is *metaphor*. These two modes of interpretation and constraints on them are examined in Chapter 6 as well.

The second class of discourse problems concerns the relation of a sentence or larger segment of text to the rest of the text of which it is a part. This is the problem of "local coherence". In Chapter 7 this problem is addressed from very much the perspective of Chapter 4 on syntax—where two segments of text are adjacent, what are the possible interpretations of this adjacency. The argument of Chapter 7 is that overwhelmingly adjacency is

interpreted as relations provided at the very foundation of the knowledge base in Chapter 5—the figure-ground relation, change of state, causality, and the similarity that allows us to build sets and other systems out of individual entities. Definitions of these *coherence relations* are given in terms provided by the knowledge base of Chapter 5. These definitions characterize what it is to recognize the relations and thereby recognize the coherence of the discourse. Examples are given of each of the coherence relations. It is shown that recognizing coherence in this way frequently leads to the solution of coreference and other discourse problems as a by-product. It is shown how this notion of coherence allows us to make very precise sense out of some of the classical and elusive concepts of discourse analysis, including "topic", "focus", "genre", and "story grammar". This account of coherence in discourse is compared with other accounts.

In Chapter 8 the perspective moves out to the environment in which the discourse occurs, and investigates how the occurrence of a discourse is to be explained as a part of ongoing events in the world. This is the problem of "global coherence" and constitutes the third set of problems that a discourse raises. Normally utterances are taken to be intentional actions in the service of a larger plan or plans the participants are engaged in. In this chapter the relation between a discourse and the plans of the participants is investigated.

Most of the book up to this point will have addressed the issue of how to make the correct interpretation of a discourse possible. But in a rich knowledge base, there will generally be many possible interpretations. Chapter 9 begins to discuss how a single "best" interpretation can be chosen for the sentence, given the various possible interpretations the theory licenses. There are two parts to this investigation—a theoretical examination and an attempt to draw lessons from practice. In the theoretical part, a framework is developed for describing optimal communication in terms of the probabilities of the predications being conveyed and the utilities of conveying them. It is shown how this unpacks at a finer grain into the scheme of weighted abduction, with a particular regime of assigning and altering weights. Then it is shown how the structured neural net representation of the weighted abduction scheme provides a biologically plausible approximation to the theoretically motivated model at the symbolic level. In the second part of Chapter 9 there is an examination of the problems that arise in interpreting the target texts and other discourse with respect to the knowledge base developed in Chapter 5.

Chapter 10 consists of detailed analyses of the target texts, to illustrate how the theory developed in this book plays out in specific instances. The

complete interpretations are shown and it is described how the solutions to the various discourse problems posed by these texts emerge from the analyses.

Finally in Chapter 11, there is a discussion of the role of a formal theory of discourse in other theoretical enterprises that are concerned with discourse, including sociology, microsociology, ethnography, psychology, and literary criticism. The probems of validating hypotheses about textual interpretations and about shared knowledge are discussed here.

Bibliography

[1] Beaugrande, Robert de, 1980. *Text, Discourse, and Process: Toward a Multidisciplinary Science of Texts*, (Advances in Discourse Processes, Vol. IV), Ablex Publishing Corporation, Norwood, New Jersey.

[2] Black, John B., 1980. "Recent Developments in Psychology with Implications for Artificial Intelligence", paper presented at First National Conference on Artificial Intelligence, Stanford, California, August 1980.

[3] Boden, Margaret, 1977. *Artificial Intelligence and Natural Man*. Basic Books, New York, New York.

[4] Chomsky, Noam, 1965. *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Massachusetts.

[5] Chomsky, Noam, 1980. *Rules and Representations*, Columbia University Press, New York, New York.

[6] Davidson, Donald, 1981. "The Material Mind", in J. Haugeland, editor, *Mind Design*, pp. 339-354, MIT Press, Cambridge, Massachusetts.

[7] Dennett, Daniel C., 1978. "Artificial Intelligence as Philosophy and as Psychology", in *Brainstorms*, pp. 109-126, Bradford Books, Cambridge, Massachusetts.

[8] Dupré, John, 1983. "The Disunity of Science", *Mind*, Vol. 92, pp.321-46.

[9] Haugeland, John, 1981. "The Nature and Plausibility of Cognitivism", in J. Haugeland, editor, *Mind Design*, pp. 243-281, MIT Press, Cambridge, Massachusetts.

[10] Lakatos, Imre, 1970. "Falsification and the Methodology of Scientific Research Programmes", in I. Lakatos and A. Musgrave, editors, *Criticism*

and the Growth of Knowledge, pp.91-196, Cambridge University Press, Cambridge, England.

[11] Mann, William C., James A. Moore, James A. Levin and James H. Carlisle, 1975. "Observation Methods for Human Dialogue", University of Southern California Information Sciences Institute Research Report 75-33, June 1975.

[12] Moore, Robert C., 1980. "Reductionism in Psychology, or Is Behaviorism Possible?", manuscript.

[13] Moravcsik, Julius M., 1980. "Chomsky's Radical Break with Modern Traditions", *The Behavioral and Brain Sciences*, Vol. 3, pp. 28-29.

[14] Newell, Allen, and Herbert A. Simon, 1976. "Computer Science as Empirical Inquiry: Symbols and Search", *Communications of the ACM*, Vol. 19, No. 3, pp. 113-126 (March 1976).

[15] Putnam, Hilary, 1981. "Reductionism and the Nature of Psychology", in J. Haugeland, editor, *Mind Design*, pp. 205-219, MIT Press, Cambridge, Massachusetts.

[16] Pylyshyn, Zenon, 1980. "Computation and Cognition: Issues in the Foundations of Cognitive Science", *The Behavioral and Brain Sciences*, Vol. 3, pp. 111-169.

[17] Pylyshyn, Zenon, 1981. "Complexity and the Study of Artificial and Human Intelligence", in J. Haugeland, editor, *Mind Design*, pp. 67-94, MIT Press, Cambridge, Massachusetts.

[18] Rubin, Andee, 1978. "A Taxonomy of Language Experiences", in *Reading: Disciplined Inquiry in Process and Practice*, The National Reading Conference, Inc., Clemson, South Carolina.

[19] Schank, Roger C., and Robert Wilensky, 1977. "Response to Dresher and Hornstein", *Cognition*, Vol. 5, No. 2, pp. 133-145.

[20] Tannen, Deborah, editor, 1982. *Spoken and Written Language: Exploring Orality and Literacy*, Ablex Publishing Corporation, Norwood, New Jersey.

[21] Van Lehn, Kurt, John Seely Brown, and James Greeno, 1983. "Competitive Argumentation in Computational Theories of Cognition", in W.

Kinsch, J. Miller, and P. Polson, editors, *Methods and Tactics in Cognitive Science*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[22] Winograd, Terry, 1977. "On Some Contested Suppositions of Generative Linguistics about the Scientific Study of Language", *Cognition*, Vol. 5, pp. 151-179.

[23] Zajonc, R. B., 1980. "Feeling and Thinking: Preferences Need No Inferences", *American Psychologist*, Vol. 35, pp. 151-175.