# DISCOURSE AND INFERENCE

**Jerry R. Hobbs**

**USC Information Sciences Institute**
**Marina del Rey, California**

**November 24, 2003**

# Chapter 3

# The Framework: Interpretation as Abduction

## 3.1 Language and Knowledge

We are able to understand language so well because we know so much. When we read the sentence

> John drove down the street in a car.

we know immediately that the driving and hence John are in the car and that the street isn't. We attach the prepositional phrase to the verb "drove" rather than to the noun "street". This is not syntactic knowledge, because in the syntactically similar sentence

> John drove down a street in Chicago.

it is the street that is in Chicago.

Therefore, a large part of the study of language should be an investigation of the question of how we use our knowledge of the world to understand discourse. This question has been examined primarily by researchers in the field of artificial intelligence (AI), in part because they have been interested in linking language with actual behavior in specific situations, which has led them to an attempt to represent and reason about fairly complex world knowledge.

In this chapter I describe how a particular kind of reasoning, called "abduction", provides a framework for addressing a broad range of problems that are posed in discourse and that require world knowledge in their solutions. As we saw in Chapter 2, first-order logic is adequate as a mode of representation for the information conveyed by sentences as well as the knowledge we bring to the discourses we interpret, but with one caveat: Reasoning must be defeasible. Here I discuss several ways that defeasible inference has been formalized in AI, and introduce abduction as one of those methods. Then in successive sections I show

- how various problems in LOCAL PRAGMATICS, such as reference resolution, metonymy, interpreting compound nominals, and word sense disambiguation can be solved via abduction;
- how this processing can be embedded in an abductive process for recognizing the syntactic structure of sentences;
- how this in turn can be embedded in a process for recognizing the structure of discourse; and
- how these can all be integrated with the recognition of the speaker's plan.

Then the specifics of the weighted abduction method and how axioms are stated in this framework are described.

To be a plausible candidate for how people process language, there must be an account of how it could be implemented in neurons. This is done in Section 3.3.11, where it is related to the SHRUTI structured connectionist model of Shastri and Ajjanagade (1993). A plausible model of language processing must also have an account of learning, and in Sections 3.3.12.1 and 3.3.12.2 such a model is presented. The first describes incremental changes to axioms to refine them, and in the second it is shown how it can be implemented in the SHRUTI model via node recruitment. The chapter closes with a comparison with relevance theory, a discussion of some of the principal outstanding research issues, and a statement of the significance of the big picture presented here.

## 3.2 Nonmonotonic Logic

The logic of mathematics is monotonic, in that once we know the truth value of a statement, nothing else we learn can change it. Virtually all commonsense knowledge beyond mathematics is uncertain or defeasible. Whatever general principles we have are usually only true most of the time or true with high probability or true unless we discover evidence to the contrary. It is almost always possible that we may have to change what we believed to be the truth value of a statement upon gaining more information. Almost all commonsense knowledge should be tagged with “insofar as I have been able to determine with my limited access to the facts and my limited resources for reasoning”. The logic of commonsense knowledge must be nonmonotonic.

The development of nonmonotonic logics has been a major focus in AI research (Ginsberg, 1987). One early attempt involved “negation as failure” (Hewitt, 1972); we assume that not $P$ is true if we fail to prove that $P$. Another early nonmonotonic logic (McDermott and Doyle, 1980) had rules of the form “If $P$ is true and $Q$ is consistent with everything else we know, then take $Q$ to be true.”

Probably the most thoroughly investigated nonmonotonic logic was that developed by McCarthy (1980). He introduced ABNORMALITY CONDITIONS which the reasoner then minimized. For example, the general fact that birds fly is expressed

$$(\forall\, x) bird(x) \;\wedge\; \neg ab_1(x) \;\supset\; fly(x)$$

That is, if $x$ is a bird and not abnormal in a way specific to this rule, then $x$ flies. Further axioms might spell out the exceptions:

$$(\forall\, x) penguin(x) \;\supset\; ab_1(x)$$

That is, penguins are abnormal in the way specific to the “birds fly” rule.

Then to draw conclusions we minimize, in some fashion, those things we take to be abnormal. If all we know about Tweety is that he is a bird, then we assume he is not abnormal, and thus we conclude he can fly. If we subsequently learn that Tweety is a penguin, we retract the assumption that he is not abnormal in that way.

A problem arises with this approach when we have many axioms with different abnormality conditions. There may be many ways to minimize the abnormalities, each leading to different conclusions. This is illustrated by an example that is known as the NIXON DIAMOND (Reiter and Criscuolo, 1981). Suppose we know that generally Quakers are pacifists. We can write this as

$$(\forall\, x) Quaker(x) \;\wedge\; \neg ab_2(x) \;\supset pacifist(x)$$

Suppose we also know that Republicans are generally not pacifists.

$$(\forall\, x) Republican(x) \;\wedge\; \neg ab_3(x) \;\supset\; \neg pacifist(x)$$

Then what do we conclude when we learn that Nixon is both a Quaker and a Republican? Assuming both abnormality conditions results in a contradiction. If we take $ab_2$ to be false, we conclude Nixon is a pacifist. If we take $ab_3$

to be false, we conclude Nixon is not a pacifist. How do we choose between the two possibilities? Researchers have made various suggestions for how to think about this problem (e.g., Shoham, 1987). In general, some scheme is needed for choosing among the possible combinations of assumptions.

In recent years there has been considerable interest in AI in the reasoning process known as abduction, or inference to the best explanation. As it is normally conceived in AI, it can be viewed as one variety of nonmonotonic logic.

## 3.3 Abduction

The simplest way to explain abduction is by comparing it with two words it rhymes with—deduction and induction. In deduction, from $P$ and $P \supset Q$, we conclude $Q$. In induction, from $P$ and $Q$, or more likely a number of instances of $P$ and $Q$ together with other considerations, we conclude $P \supset Q$. Abduction is the third possibility. From an observable $Q$ and a general principle $P \supset Q$, we conclude that $P$ must be the underlying reason that $Q$ is true. We assume $P$ because it explains $Q$.

Of course, there may be many such possible $P$'s, some contradicting others, and therefore any method of abduction must include a method for evaluating and choosing among alternatives. At a first cut, suppose in trying to explain $Q$ we know $P \wedge R \supset Q$ and we know $R$. Then $R$ provides partial evidence that $Q$ is true, making the assumption of $P$ more reasonable. In addition, if we are seeking to explain two things, $Q_1$ and $Q_2$, then it is reasonable to favor assuming a $P$ that explains both of them rather than a different explanation for each.

The conclusions we draw in this way are only assumptions and may have to be retracted later if we acquire new, contradictory information. That is, this method of reasoning is nonmonotonic.

Abduction has a history. Prior to the late seventeenth century science was viewed as deductive, at least in the ideal. It was felt that, on the model of Euclidean geometry, one should begin with propositions that were self-evident and deduce whatever consequences one could from them. As was articulated in Chapter 1, the modern view of scientific theories (Lakatos, 1970), is quite different. One tries to construct abstract theories from which observable events can be deduced or predicted. There is no need for the abstract theories to be self-evident, and they usually are not. It is only necessary for them to predict as broad a range as possible of the observable data and for them to be "elegant", whatever that means. Thus, the modern view is that science is fundamentally abductive. We seek hidden principles or causes from which we can deduce the observable evidence.

This view of science, and hence the notion of abduction, can be seen first, insofar as I am aware, in some passages in Newton's *Principia* (1934 [1686]). At the end of *Principia*, in a justification for not seeking the cause of gravity, he says, "And to us it is enough that gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea." (Newton 1934:547) The justification for gravity ($P$) and its laws ($P \supset Q$) is not in their self-evidential nature but in what they account for ($Q$).

In the eighteenth century, the German philosopher Christian Wolff (1963 [1728]) shows, to my knowledge, the earliest *explicit* awareness of the importance of abductive reasoning. He presents almost the standard Euclidean account of certain knowledge, but with an important provision in his recognition of the inevitability and importance of hypotheses:

> Philosophy must use hypotheses insofar as they pave the way to the discovery of certain truth. For in a philosophical hypothesis certain things which are not firmly established are assumed because they provide a reason for things which are observed to occur. Now if we can also deduce other things which are not observed to occur, then we have the opportunity to either observe or experimentally detect things which otherwise we might not have noticed. In this way we become more certain as to whether or not anything contrary to experience follows from the hypothesis. If we deduce things which are contrary to experience, then the hypothesis is false. If the deductions agree with experience, then the probability of the hypothesis is increased. And thus the way is paved for the discovery of certain truth. (Wolff 1963:67)

He also recognizes the principle of parsimony: "If one cannot necessarily deduce from a hypothesis the things for which it is assumed, then the hypothesis is spurious." (Wolff, 1963:68) However, he views hypotheses as only provisional, awaiting deductive proof.

The term "abduction" was first used by C. S. Pierce (e.g., 1955). His definition of it is as follows:

> The surprising fact, Q, is observed;
> But if P were true, Q would be a matter of course,
> Hence, there is reason to suspect that P is true. (Pierce, 1955:151)

(He actually used A and C for P and Q.) Pierce says that “in pure abduction, it can never be justifiable to accept the hypothesis otherwise than as an interrogation”, and that “the whole question of what one out of a number of possible hypotheses ought to be entertained becomes purely a question of economy.” That is, there must be an evaluation scheme for choosing among possible abductive inferences.

The earliest formulation of abduction in artificial intelligence was by Morgan (1971). He showed how a complete set of truth-preserving rules for generating theorems could be turned into a complete set of falsehood-preserving rules for generating hypotheses.

The first use of abduction in an AI application was by Pople (1973), in the context of medical diagnosis. He gave the formulation of abduction sketched above and showed how it can be implemented in a theorem-proving framework. Literals (or propositions) that are “abandoned by deduction in the sense that they fail to have successor nodes” (Pople, 1973:150) are taken as the candidate hypotheses. That is, one tries to prove the symptoms and signs exhibited and the parts of a potential proof that cannot be proven are the candidate hypotheses. Those hypotheses are best that account for the most data, and in service of this principle, he introduced factoring or synthesis, which attempts to unify goal literals. Hypotheses where this is used are favored. That is, that explanation is best that minimizes the number of causes.

Work on abduction in artificial intelligence was revived in the 1980s at several sites. Reggia and his colleagues (e.g., Reggia et al., 1983; Reggia, 1985) formulated abductive inference in terms of parsimonious covering theory. Charniak and McDermott (1985) presented the basic pattern of abduction and then discussed many of the issues involved in trying to decide among alternative hypotheses on probabilistic grounds. Cox and Pietrzykowski (1986) present a formulation in a theorem-proving framework that is very similar to Pople’s, though apparently independent. It is especially valuable in that it considers abduction abstractly, as a mechanism with a variety of possible applications, and not just as a handmaiden to diagnosis.

Josephson and Josephson (1994) provide a comprehensive treatment of abduction, its philosophical background, its computational properties, and its utilization in AI applications.

I have indicated that the practice of science is fundamentally abductive. The extension of abduction to ordinary cognitive tasks is very much in line with the popular view in cognitive science that people going about in the world trying to understand it are scientists in the small. This view can be extended to natural language understanding—interpreting discourse is coming up with the best explanation for what is said.

The first appeal to something like abduction that I am aware of in natural language understanding was by Grice (1967, 1989), when he introduced the notion of *conversational implicature* to handle examples like the following:

A: How is John doing on his new job at the bank?
B: Quite well. He likes his colleagues and he hasn't embezzled any money yet.

Grice argues that in order to see this as coherent, we must assume, or draw as a conversational implicature, that both A and B know that John is dishonest. Although he does not say so, an implicature can be viewed as an abductive move for the sake of achieving the best interpretation.

Lewis (1979) introduces the notion of *accommodation* in conversation to explain the phenomenon that occurs when you "say something that requires a missing presupposition, and straightaway that presupposition springs into existence, making what you said acceptable after all." The hearer accommodates the speaker.

Thomason (1990) argued that Grice's conversational implicatures are based on Lewis's rule of accommodation. We might say that implicature is a procedural characterization of something that, at the functional or interactional level, appears as accommodation. Implicature is the way we do accommodation.

In the middle 1980s researchers at several sites began to apply abduction to natural language understanding (Norvig, 1983, 1987; Wilensky, 1983; Wilensky et al., 1988; Charniak and Goldman, 1988, 1989; Hobbs et al., 1988; Hobbs et al., 1993). At least in the last case the recognition that implicature was a use of abduction was a key observation in the development of the framework.

Norvig, Wilensky, and their associates proposed an operation called *concretion*, one of many that take place in the processing of a text. It is a "kind of inference in which a more specific interpretation of an utterance is made than can be sustained on a strictly logical basis" (Wilensky et al., 1988:50). Thus, "to use a pencil" generally means to write with a pencil, even though one could use a pencil for many other purposes.

Charniak and his associates also developed an abductive approach to interpretation. Charniak (1986) expressed the fundamental insight: "A standard platitude is that understanding something is relating it to what one already knows. ... One extreme example would be to prove that what one is told must be true on the basis of what one already knows. ... We want to prove what one is told *given certain assumptions*." (Charniak, 1986:585)

Charniak and Goldman developed an interpretation procedure that incrementally built a belief network (Pearl, 1988), where the links between the nodes, representing influences between events, were determined from axioms expressing world knowledge. They felt that one could make not unreasonable estimates of the required probabilities, giving a principled semantics to the numbers. The networks were then evaluated and ambiguities were resolved by looking for the highest resultant probabilities.

Stickel invented a method called *weighted abduction* (Stickel 1988; Hobbs et al., 1993) that builds the evaluation criteria into the proof process. Briefly, propositions to be proved are given an assumption cost—what you will have to pay to assume them. When we backchain over a rule of the form $P \supset Q$, the cost is passed back from $Q$ to $P$, according to a weight associated with $P$. Generally, $P$ will cost more to assume than $Q$, so that short proofs are favored over long ones. But if partial evidence is found, for example, if $P \wedge R \supset Q$ and we can prove $R$, then it will cost less to assume $P$ than to assume $Q$, and we get a more specific interpretation. In addition, if we need to prove $Q_1$ and $Q_2$ and $P$ implies both, then it will cost less to assume $P$ than to assume $Q_1$ and $Q_2$. This feature of the method allows us to exploit the implicit redundancy inherent in natural language discourse.

Weighted abduction suggests a simple way to incorporate the uncertainty of knowledge into the axioms expressing the knowledge. Propositions can be assumed at a cost. Therefore, we can have propositions whose only role is to be assumed and to levy a cost. For example, let's return to the rule that birds fly. We can express it with the axiom

$$(\forall\, x)[bird(x) \wedge etc_1(x) \supset fly(x)]$$

That is, if $x$ is a bird and some other unspecified conditions hold for $x$ ($etc_1(x)$), then $x$ flies. The predicate $etc_1$ encodes the unspecified conditions. There will never be a way to prove it; it can only be assumed at cost. The cost of $etc_1$ will depend inversely on the certainty of the rule that birds fly. It will cost to use this rule, but the lowest-cost proof of everything we are trying to explain may nevertheless involve this rule and hence the inference that birds fly. We know that penguins don't fly:

$$(\forall\, x)[penguin(x) \supset \neg fly(x)]$$

If we know Tweety is a penguin, we know he doesn't fly. Thus, to assume $etc_1$ is true of Tweety would lead to a contradiction, so we don't. The relation between the *etc* predicates and the abnormality predicates of McCarthy's nonmonotonic logic is obvious: $etc_1$ is just $\neg ab_1$.

The framework of "Interpretation as Abduction" (IA) (Hobbs et al. 1993) follows directly from this method of abductive inference, and it is the IA framework that is used in the rest of this book. Whereas in Norvig and Wilensky's work, abduction or concretion was one process among many involved in natural language understanding, in the IA framework abduction is the whole story. Whereas in Charniak and Goldman's work, specific procedures involving abduction are implemented to solve specific interpretation problems, in the IA framework there is only one procedure—abduction—that is used to explain or prove the logical form of the text, and the solutions to specific interpretation problems fall out as byproducts of this process.

It should be pointed out that in addition to what is presented below there have been a number of other researchers who have used abduction for various natural language understanding problems, including Nagao (1989) for resolving syntactic ambiguity, Dasigi (1988) for resolving lexical ambiguity, Rayner (1993) for asking questions of a database, Ng and Mooney (1990) and Lascarides and Oberlander (1992) for recognizing discourse structure, McRoy and Hirst (1991) for making repairs in presupposition errors, Appelt and Pollack (1990) for recognizing the speaker's plan, and Harabagiu and Moldovan (1998, 2002) for general text understanding and question-answering using WordNet as a knowledge base.

## 3.4 Interpretation as Abduction

In the IA framework we can describe very concisely what it is to interpret a sentence:

Prove the logical form of the sentence,
together with the selectional constraints that predicates
impose on
their arguments,
allowing for coercions,
Merging redundancies where possible,
Making assumptions where necessary.

By the first line we mean "prove, or derive in the logical sense, from the predicate calculus axioms in the knowledge base, the logical form that has been produced by syntactic analysis and semantic translation of the sentence."

We can view this as in Figure 3.1. In a discourse situation, the speaker and hearer both have their sets of private beliefs, and there is a large overlapping set of mutual beliefs. An utterance lives on the boundary between mutual belief and the speaker's private beliefs. It is a bid to extend the area of mutual belief to include some private beliefs of the speaker's. It is anchored referentially in mutual belief, and when we succeed in proving the logical form and the constraints, we are recognizing this referential anchor. This is the given information, the definite, the presupposed. Where it is necessary to make assumptions, the information comes from the speaker's private beliefs, and hence is the new information, the indefinite, the asserted. Merging redundancies is a way of getting a minimal, and hence a best, interpretation.

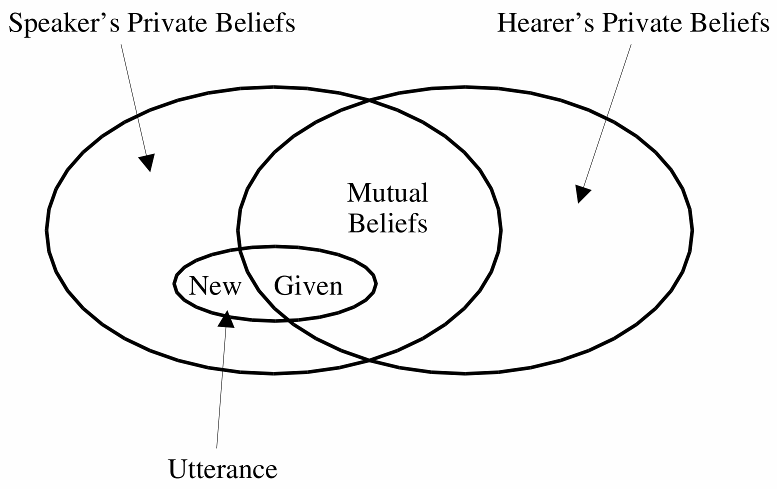

Figure 3.1: The Discourse Situation

Merging redundancies and minimizing the assumptions result naturally from the method of weighted abduction.

## 3.5 Abduction and Local Pragmatics

Local pragmatics encompasses those problems that are posed within the scope of individual sentences, even though their solution will generally require greater context and world knowledge. Included under this label are the resolution of coreference, resolving syntactic and lexical ambiguity, interpreting metonymy and metaphor, and finding specific meanings for vague predicates such as in the compound nominal.

Consider a simple example that contains three of these problems.

The Boston office called.

This sentence poses at least three local pragmatics problems, the problems of resolving the reference of "the Boston office", expanding the metonymy to "[Some person at] the Boston office called", and determining the implicit relation between Boston and the office. Let us put these problems aside for the moment, however, and interpret the sentence according to the IA characterization. We must prove abductively the logical form of the sentence together with the constraint "call" imposes on its agent, allowing for a coercion. That is, we must prove abductively the expression (ignoring tense and some other complexities)

$$(\exists\, x, y, z, e) call'(e, x) \wedge person(x) \wedge rel(x, y) \wedge office(y) \wedge Boston(z) \\ \wedge\, nn(z, y)$$

That is, there is a calling event $e$ by $x$ where $x$ is a person. $x$ may or may not be the same as the explicit subject of the sentence, but it is at least related to it, or coercible from it, represented by $rel(x, y)$. $y$ is an office and it bears some unspecified relation $nn$ to $z$ which is Boston. $person(x)$ is the requirement that $call'$ imposes on its agent $x$. The predicate $rel$ is for accomodating metonymy. How it is introduced is discussed in the next section. The interesting and important question of what specific relations can instantiate it is dealt with in Chapter 6.

The sentence can be interpreted with respect to a knowledge base of mutual knowledge that contains the following facts:

$$Boston(B_1)$$

that is, $B_1$ is the city of Boston.

$$office(O_1)\ \wedge\ in(O_1, B_1)$$

that is, $O_1$ is an office and is in Boston.

$$person(J_1)$$

that is, John $J_1$ is a person.

$$work\text{-}for(J_1, O_1)$$

that is, John $J_1$ works for the office $O_1$.

$$(\forall y,z) in(y,z) \supset nn(z,y)$$

that is, if $y$ is in $z$, then $z$ and $y$ are in a possible compound nominal relation.

$$(\forall x,y) work\text{-}for(x,y) \supset rel(x,y)$$

that is, if $x$ works for $y$, then $y$ can be coerced into $x$.

Given these axioms, the proof of all of the logical form is straightforward except for the conjunct $call'(e,x)$. Hence, we assume that; it is the new information conveyed by the sentence.

This interpretation is illustrated in the proof graph of Figure 1, where a rectangle is drawn around the assumed literal $call'(e,x)$. Such proof graphs play the same role in interpretation as parse trees play in syntactic analysis. They are pictures of the interpretations, and we will see many such diagrams in this book.

Now notice that the three local pragmatics problems have been solved as a by-product. We have resolved "the Boston office" to $O_1$. We have determined the implicit relation in the compound nominal to be *in*. And we have expanded the metonymy to "John, who works for the Boston office, called."

For an illustration of the resolution of lexical ambiguity, consider an example from Hirst (1987):

> The plane taxied to the terminal.

The words "plane", "taxied", and "terminal" are all ambiguous.

Suppose the knowledge base consists of the following axioms:

$$(\forall x) airplane(x) \supset plane(x)$$

or an airplane is a plane.

$$(\forall x) wood\text{-}smoother(x) \supset plane(x)$$

or a wood smoother is a plane.

$$(\forall x,y) move\text{-}on\text{-}ground(x,y) \wedge airplane(x) \supset taxi(x,y)$$

or for an airplane $x$ to move on the ground to $y$ is for it to taxi to $y$.

$$(\forall x,y) ride\text{-}in\text{-}cab(x,y) \wedge person(x) \supset taxi(x,y)$$

or for a person $x$ to ride in a cab to $y$ is for $x$ to taxi to $y$.

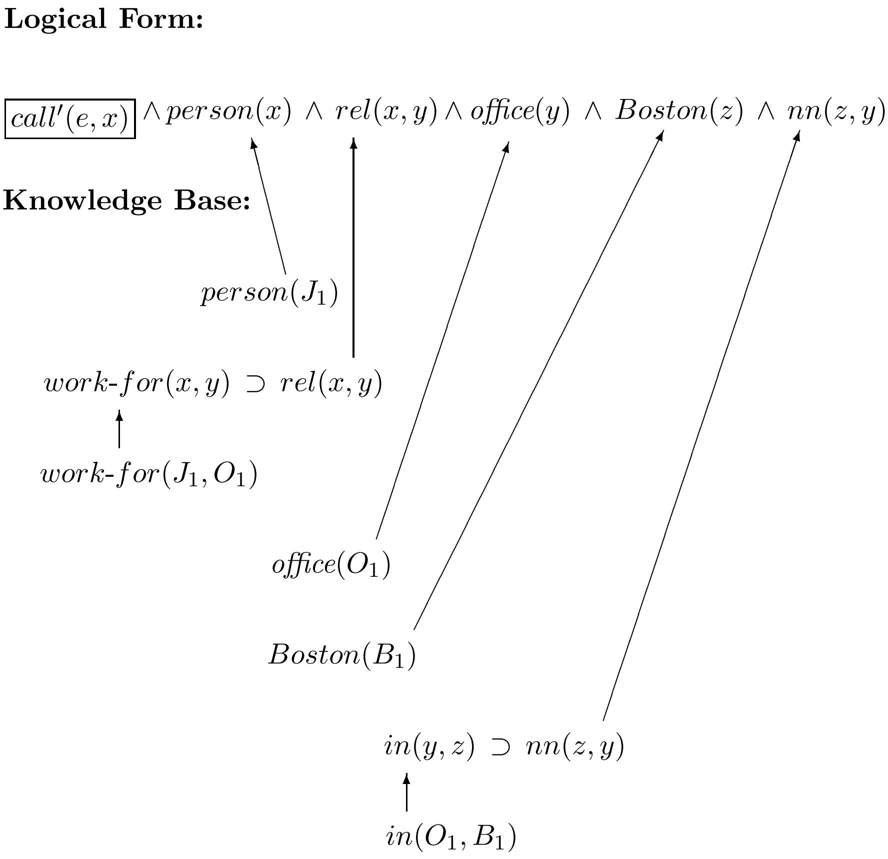
**Logical Form:**
$\boxed{call'(e,x)} \wedge person(x) \wedge rel(x,y) \wedge office(y) \wedge Boston(z) \wedge nn(z,y)$
**Knowledge Base:**
$person(J_1)$
$work\text{-}for(x,y) \supset rel(x,y)$
$work\text{-}for(J_1,O_1)$
$office(O_1)$
$Boston(B_1)$
$in(y,z) \supset nn(z,y)$
$in(O_1,B_1)$

Figure 3.2: Interpretation of “The Boston office called.”

$$(\forall\, y)airport\text{-}terminal(y) \supset terminal(y)$$

or an airport terminal is a terminal.

$$(\forall\, y)computer\text{-}terminal(y) \supset terminal(y)$$

or a computer terminal is a terminal.

$$(\forall\, z)airport(z) \supset (\exists\, x,y)airplane(x) \wedge airport\text{-}terminal(y)$$

or airports have airplanes and airport terminals.

The logical form of the sentence will be, roughly,

$$(\exists\, x,y)plane(x) \wedge taxi(x,y) \wedge terminal(y)$$

The minimal proof of this logical form will involve assuming the existence of an airport, deriving from that the airplane, and thus the plane, and the airport terminal, and thus the terminal, assuming $x$ is moving on the ground to $y$, and recognizing the redundancy of the airplane with the one in that reading of "taxi". This interpretation is illustrated in Figure 2.

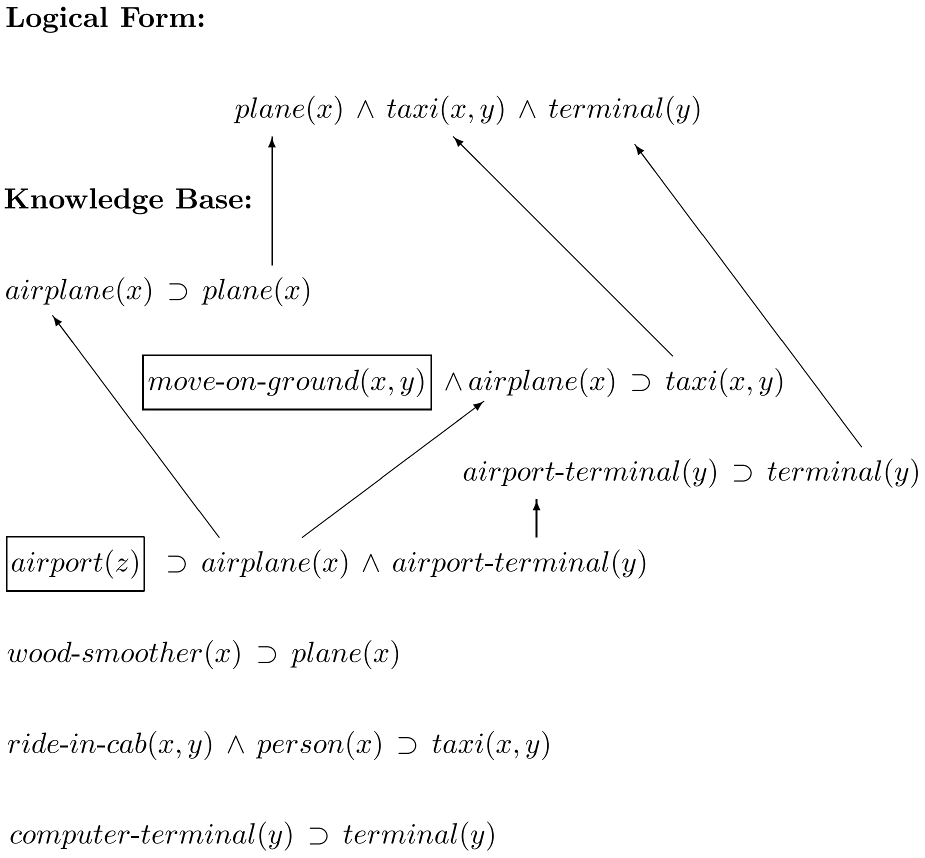
**Logical Form:**
$plane(x) \wedge taxi(x,y) \wedge terminal(y)$
**Knowledge Base:**
$airplane(x) \supset plane(x)$
$\boxed{move\text{-}on\text{-}ground(x,y)} \wedge airplane(x) \supset taxi(x,y)$
$airport\text{-}terminal(y) \supset terminal(y)$
$\boxed{airport(z)} \supset airplane(x) \wedge airport\text{-}terminal(y)$
$wood\text{-}smoother(x) \supset plane(x)$
$ride\text{-}in\text{-}cab(x,y) \wedge person(x) \supset taxi(x,y)$
$computer\text{-}terminal(y) \supset terminal(y)$

Figure 3.3: Interpretation of "The plane taxied to the terminal."

Another possible interpretation would be one in which we assumed that a wood smoother, a ride in a taxi, and a computer terminal all existed. It is because weighted abduction favors merging redundancies that the correct interpretation is the one chosen. That interpretation allows us to minimize the assumptions we make.

## 3.6 Syntax by Abduction

In Chapter 4 an extensive subset of English grammar is described in detail, largely following Pollard and Sag's (1994) Head-Driven Phrase Structure Grammar but cast into the IA framework. In this treatment, the predicate *Syn* is used to express the relation between a string of words and the eventuality it conveys. Certain axioms involving *Syn*, the *composition axioms*, describe how the eventuality conveyed emerges from the concatenation of strings. Other axioms, the *lexical axioms*, link *Syn* predications about words with the corresponding logical-form fragments. There are also *alternation axioms* which alter the places in the string of words where predicates find their arguments.

In this chapter, a simplified version of the predicate *Syn* will be used. We will take *Syn* to be a predicate of seven arguments.

$$Syn(w, e, f, x, a, y, b)$$

$w$ is a string of words. $e$ is the eventuality described by this string. $f$ is the category of the head of the phrase $w$. If the string $w$ contains the logical subject of the head, then the arguments $x$ and $a$ are the empty symbol "$-$". Otherwise, $x$ is a variable refering to the logical subject and $a$ is its category. Similarly, $y$ is either the empty symbol or a variable refering to the logical object and $b$ is either the empty symbol or the category of the logical object. For example,

$$Syn(\text{“reads a novel”}, e, \mathbf{v}, x, \mathbf{n}, -, -)$$

says that the string of words "reads a novel" is a phrase describing an eventuality $e$ and has a head of category **verb**. Its logical object "a novel" is in the string itself, so the last two arguments are the empty symbol. Its logical subject is not part of the string, so the fourth argument is the variable $x$ standing for the logical subject and the fifth argument specifies that the phrase describing it must have a noun as its head. In Chapter 4 the full *Syn* predicate contains argument positions for further complements and filler-gap information, and the category arguments can record syntactic features as well.

Two of the most important composition axioms are the following:

$$(\forall\, w_1, w_2, x, a, e, f) Syn(w_1, x, a, -, -, -, -) \wedge Syn(w_2, e, f, x, a, -, -)$$
$$\supset\ Syn(w_1 w_2, e, f, -, -, -, -)$$
$$(\forall\, w_1, w_2, e, f, x, a, y, b) Syn(w_1, e, f, x, a, y, b) \wedge Syn(w_2, y, b, -, -, -, -)$$
$$\supset\ Syn(w_1 w_2, e, f, x, a, -, -)$$

The first axiom corresponds to the traditional "S → NP VP" rule. It says that if $w_1$ is a string describing an entity $x$ and headed by a word of category $a$ and $w_2$ is a string describing eventuality $e$, headed by a word of category $f$, and lacking a logical subject $x$ of category $a$, then the concatenation $w_1w_2$ is a string describing eventuality $e$ and headed by a word of category $f$.

The second axiom corresponds to the traditional "VP → V NP" rule. It says that if $w_1$ is a string describing eventuality $e$, headed by a word of category $f$, and lacking a logical subject $x$ of category $a$ and a logical object $y$ of category $b$ and $w_2$ is a string describing an entity $y$ and headed by a word of category $b$, then the concatenation $w_1w_2$ is a string describing eventuality $e$, headed by a word of category $f$, and lacking a logical subject $x$ of category $a$, but not lacking a logical object.

A typical lexical axiom is the following:

$$\begin{aligned}&(\forall\, e, x, y) past(e) \wedge read'(e, x, y) \wedge person(x) \wedge text(y)\\ &\quad \supset Syn(\text{"read"}, e, \mathbf{v}, x, \mathbf{n}, y, \mathbf{n})\end{aligned}$$

That is, if $e$ is the eventuality in the past of a person $x$ reading a text $y$, then the verb "read" can be used to describe $e$ provided noun phrases describing $x$ and $y$ are found in the appropriate places, as specified by composition axioms. Lexical axioms thus encode the logical form fragment corresponding to a word ($past(e) \wedge read'(e, x, y)$), selectional constraints ($person(x)$ and $text(y)$), the spelling (or in a more detailed account, the phonology) of the word ("read"), its category (verb), and the syntactic constraints on its complements (that $x$ and $y$ must come from noun phrases). The lexical axioms constitute the interface between syntax and world knowledge; knowledge about reading is encoded in axioms involving the predicate $read'$, whereas knowledge of syntax is encoded in axioms involving $Syn$, and these two are linked here.

Interpreting a sentence $W$ is then proving the expression

$$(\exists\, e) Syn(W, e, \mathbf{v}, -, -, -, -)$$

i.e., proving that $W$ is headed by a verb, describes some eventuality $e$, and is complete in that it does not lack a logical subject and logical object. The parse of the sentence is found because composition axioms are used in the proof. The logical form is generated because that part of the proof bottoms out in lexical axioms. The local pragmatics problems are solved because that logical form is then proved. That is, in the course of proving that a string of words is a grammatical, interpretable sentence, the interpretation process backchains through composition axioms to lexical axioms (the syntactic processing) and then is left with the logical form of the sentence to be proved. A proof of this logical form was the IA characterization of the interpretation of a sentence given in the previous section.

The proof graph of the syntactic part of the interpretation of “John read *Ulysses*” is shown in Figure 3.4. Note that knowledge that John is a person and *Ulysses* is a text is used to establish the selectional constraints associated with “read”.

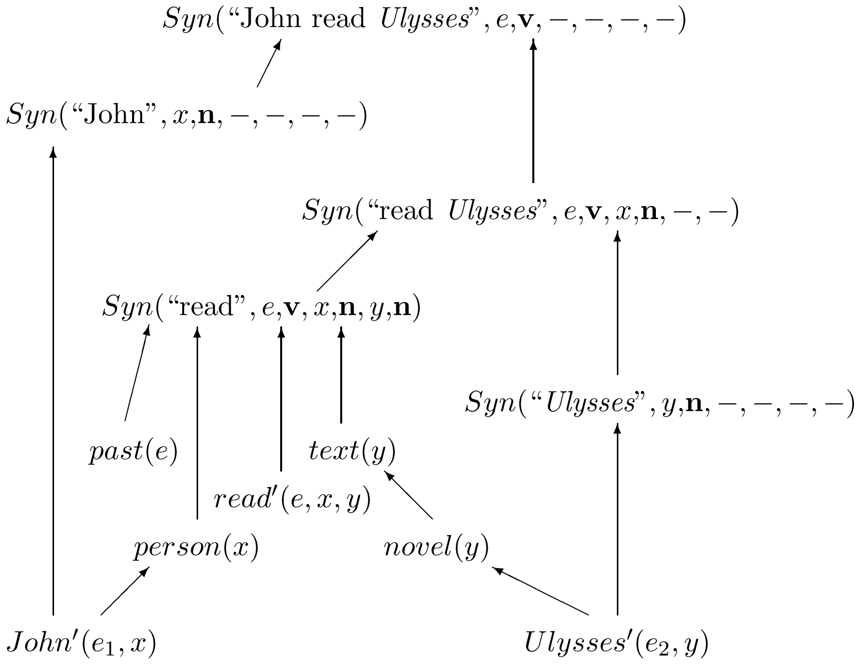
Syn(“John read Ulysses”, e,v, –, –, –, –)
Syn(“John”, x,n, –, –, –, –)
Syn(“read Ulysses”, e,v, x,n, –, –)
Syn(“read”, e,v, x,n, y,n)
Syn(“Ulysses”, y,n, –, –, –, –)
past(e)
text(y)
read′(e, x, y)
person(x)
novel(y)
John′(e1, x)
Ulysses′(e2, y)

Figure 3.4: Parse of “John read *Ulysses*.”

In Chapter 4 there are about a dozen composition axioms, corresponding to similar rules in Pollard and Sag (1994). There is one lexical axiom for every word sense and subcategorization pattern; the lexical axioms constitute the lexicon.

There are also a number of alternation axioms that handle such things as passive constructions. These axioms alter the order of, or otherwise modify, the arguments of the predicate associated with a construction’s head. Metonymic coercion relations can be introduced by means of an alternation axiom of the form

$$Syn(w, e, f, x_0, a, y, b) \wedge rel(x_0, x) \supset Syn(w, e, f, x, a, y, b)$$

That is, a word or phrase $w$ looking for a subject referring to $x_0$ can be used to describe the same situation if its subject refers to $x$ instead, where $x$ is related to $x_0$ by a coercion relation $rel$.

As presented so far, abduction plays no role in this encoding of syntactic knowledge. Syntactic processing is just logical deduction. The principal advantage of the framework is that it allows syntactic analysis to be done with other interpretion processes in a uniform framework. In addition, various sorts of ungrammaticality—telegraphic discourse, disfluencies, scrambling—can be handled by means of assumptions.

In this section we have recast the problem of interpreting a sentence as one of proving that the string of words is a grammatical, interpretable sentence. Local pragmatics is subsumed under that characterization in the word "interpretable". It is well known that there are interactions between syntactic processing and pragmatics. By solving both problems with one proof and choosing among proofs by means of a common evaluation metric, we can model those interactions. Sometimes less favored solutions will be chosen in each part of the proof because that results in the lowest-cost proof overall.

In the next section we will see how this picture can be embedded in an even larger picture.

## 3.7 Recognizing Discourse Structure

When two segments of discourse are adjacent, that very adjacency conveys information. Each segment, insofar as it is coherent, conveys information about a situation or eventuality, and the adjacency of the segments conveys the suggestion that the two situations are related in some fashion, or are parts of larger units that are related. Part of what it is to understand a discourse is to discover what that relation is.

Overwhelmingly, the relations that obtain between discourse segments are based on causal, similarity, or figure-ground relations between the situations they convey. We can thus define a number of *coherence relations* in terms of the relations between the situations. This aspect of discourse structure can be built into the abduction framework.

Suppose $w_1$ and $w_2$ are two adjacent segments of discourse and that $w_1w_2$ is their concatenation. If $Segment(w, e)$ says that the string $w$ is a coherent segment of discourse describing the eventuality $e$ and $CoherenceRel(e_1, e_2, e)$ says that there is a coherence relation between $e_1$ and $e_2$ and that the combination of the two conveys $e$, then we can express the basic composition rule for discourse as

$$\begin{aligned}&(\forall\, w_1, w_2, e_1, e_2, e) Segment(w_1, e_1) \wedge Segment(w_2, e_2)\\ &\qquad \wedge CoherenceRel(e_1, e_2, e)\\ &\quad \supset Segment(w_1 w_2, e)\end{aligned}$$

That is, when we combine two coherent segments of discourse with a coherence relation we get a coherent segment of discourse. By applying this successively to a stretch of discourse, we get a tree-like structure for the whole discourse.

This process bottoms out in sentences, after which syntactic rules tell us the structure and meaning of the string of words. This is captured by the rule

$$(\forall\, w, e) Syn(w, e, \mathbf{v}, -, -, -, -) \supset Segment(w, e)$$

That is, a grammatical sentence conveying $e$ is a coherent segment of discourse conveying $e$.

In the previous section the solution to local pragmatics problems—proving the logical form—was embedded in the problem of finding the syntactic structure of, or parsing, a sentence. These two axioms now embed parsing the sentence in the problem of recognizing the discourse structure of the whole text. If $W$ is a text, then interpreting $W$ is a matter of proving that it is a coherent segment of discourse conveying some eventuality $e$:

$$(\exists\, e) Segment(W, e)$$

Now let us consider an example. Explanation is a coherence relation, and a first approximation of a definition of the Explanation relation would be that the eventuality described by the second segment causes the eventuality described by the first:

$$(\forall\, e_1, e_2) cause(e_2, e_1) \supset CoherenceRel(e_1, e_2, e_1)$$

That is, if what is described by the second segment could cause what is described by the first segment, then there is a coherence relation between the segments. In explanations, what is explained is the dominant segment (the *nucleus* in the terms of Rhetorical Structure Theory (Mann and Thompson, 1986)), so it is $e_1$ that is described by the composed segment. Hence, the third argument of *CoherenceRel* is $e_1$.

Consider a variation on a classic example of pronoun resolution difficulties from Winograd (1972):

> The police prohibited the women from demonstrating.
> They feared violence.

How do we know “they” in the second sentence refers to the police and not to the women?

As in Section 6, we will ignore this local pragmatics problem and proceed to interpret the text by abduction. To interpret the text is to prove abductively the expression

$$(\exists\, e)Segment(\text{“The police \ldots violence.”},\ e)$$

This involves proving that each sentence is a segment, by proving they are grammatical, interpretable sentences, and proving there is a coherence relation between them. To prove they are sentences, we would tap into an expanded version of the sentence grammar of Section 3.3.6. This would bottom out in the logical forms of the sentences, via the lexical axioms, and thus require us to prove abductively those logical forms.

One way to prove there is a coherence relation between the sentences is to prove there is an Explanation relation between them by showing there is a causal relation between the eventualities they describe.

After back-chaining in this manner, we are faced with proving the expression

$$\begin{aligned}&(\exists\, e_1, p, d, w, e_2, y, v, z) police(p) \wedge prohibit'(e_1, p, d) \wedge demonstrate'(d, w)\\ &\qquad \wedge cause(e_2, e_1) \ \wedge\ fear'(e_2, y, v) \ \wedge\ violent'(v, z)\end{aligned}$$

That is, there is a prohibiting event $e_1$ by the police $p$ of a demonstrating event $d$ by the women $w$. There is a fearing event $e_2$ by someone $y$ (“they”) of violence $v$ by someone $z$. The fearing event $e_2$ causes the prohibiting event $e_1$. This expression is just the (simplified) logical forms of the two sentences, plus the hypothesized causal relation between them.

Suppose, plausibly enough, we have in our knowledge base the following axioms:

$$(\forall\, e_2, y, v) fear'(e_2, y, v) \ \supset\ (\exists\, d_2) diswant'(d_2, y, v) \wedge cause(e_2, d_2)$$

That is, if $e_2$ is a fearing by $y$ of $v$, then that will cause the state $d_2$ of $y$ not wanting or “diswanting” $v$.

$$(\forall\, d, w) demonstrate'(d, w) \ \supset\ (\exists\, v, z) cause(d, v) \ \wedge\ violent'(v, z)$$

That is, demonstrations cause violence.

$$\begin{aligned}&(\forall\, d, v, d_2, y) cause(d, v) \ \wedge\ diswant'(d_2, y, v)\\ &\qquad \supset\ (\exists\, d_1) diswant'(d_1, y, d) \ \wedge\ cause(d_2, d_1)\end{aligned}$$

That is, if someone $y$ diswants $v$ and $v$ is caused by $d$, then that will cause $y$ to diswant $d$ as well. If you don't want the effect, you don't want the cause.

$$\begin{aligned}&(\forall\, d_1, p, d) diswant'(d_1, p, d) \,\wedge\, authority(p)\\ &\quad\supset\, (\exists\, e_1) prohibit'(e_1, p, d) \,\wedge\, cause(d_1, e_1)\end{aligned}$$

That is, if those in authority diswant something, that will cause them to prohibit it.

$$(\forall\, e_1, e_2, e_3) cause(e_1, e_2) \,\wedge\, cause(e_2, e_3) \,\supset\, cause(e_1, e_3)$$

That is, *cause* is transitive.

$$(\forall\, p) police(p) \,\supset\, authority(p)$$

That is, the police are in authority.

From these axioms, we can prove all of the above logical form except the propositions $police(p)$, $demonstrate'(d, w)$, and $fear'(f, y, v)$, which we assume. This is illustrated in Figure 3. Notice that in the course of doing the proof, we unify $y$ with $p$, thus resolving the problematic pronoun reference that originally motivated this example. "They" refers to the police.

One can imagine a number of variations on this example. If we had not included the axiom that demonstrations cause violence, we would have had to assume the violence and the causal relation between demonstrations and violence. Moreover, other coherence relations might be imagined here by constructing the surrounding context in the right way. It could be followed by the sentence "But since they had never demonstrated before, they did not know that violence might result." In this case, the second sentence would play a subordinate role to the third, forcing the resolution of "they" to the women. Each example, of course, has to be analyzed on its own, and changing the example changes the analysis. In Winograd's original version of this example,

> The police prohibited the women from demonstrating, because they feared violence.

the causality was explicit, thus eliminating the coherence relation as a source of ambiguity. The literal $cause(e_2, e_1)$ would be part of the logical form.

Winograd's contrasting text, in which "they" is resolved to the women, is

> The police prohibited the women from demonstrating, because they advocated violence.

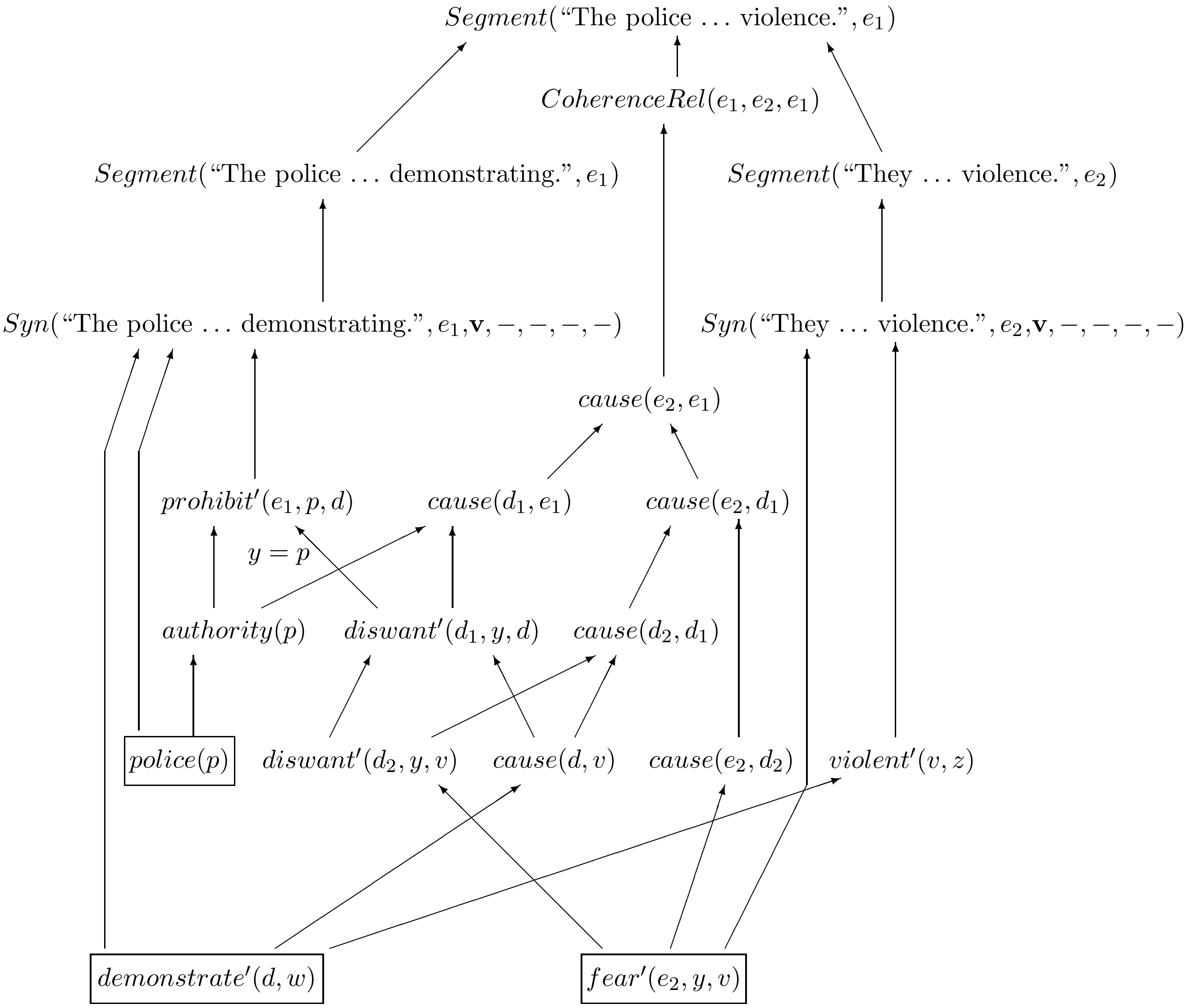
$Segment(\text{“The police … violence.”}, e_1)$
$CoherenceRel(e_1, e_2, e_1)$
$Segment(\text{“The police … demonstrating.”}, e_1)$
$Segment(\text{“They … violence.”}, e_2)$
$Syn(\text{“The police … demonstrating.”}, e_1, \mathbf{v}, -, -, -, -)$
$Syn(\text{“They … violence.”}, e_2, \mathbf{v}, -, -, -, -)$
$cause(e_2, e_1)$
$prohibit'(e_1, p, d)$
$cause(d_1, e_1)$
$cause(e_2, d_1)$
$y = p$
$authority(p)$
$diswant'(d_1, y, d)$
$cause(d_2, d_1)$
$police(p)$
$diswant'(d_2, y, v)$
$cause(d, v)$
$cause(e_2, d_2)$
$violent'(v, z)$
$demonstrate'(d, w)$
$fear'(e_2, y, v)$

Figure 3.5: Interpretation of “The police prohibited the women from demonstrating. They feared violence.”

Here we would need the facts that when one demonstrates one advocates and that advocating something tends to bring it about. Then showing a causal relation between the clauses will result in "they" being identified with the demonstrators.

## 3.8 Recognizing the Speaker's Plan

As presented so far, understanding discourse is seeing the world of the text as coherent, which in turn involves viewing the content of the text as observables to be explained. The focus has been on the information conveyed explicitly or implicitly by the discourse. We can call this the *informational* account of a discourse.

But utterances are embedded in the world as well. They are produced to realize a speaker's intention, or more generally, they are actions in the execution of a speaker's plan to achieve some goal. The description of how a discourse realizes the speakers' goals may be called the *intentional* account of the discourse.

Let us consider the intentional account from the broadest perspective. An intelligent agent is embedded in the world and must, at each instant, understand the current situation. The agent does so by finding an explanation for what is perceived. Put differently, the agent must explain why the complete set of observables encountered constitutes a coherent situation. Other agents in the environment are viewed as intentional, that is, as planning mechanisms, and this means that the best explanation of their observable actions is most likely to be that the actions are steps in a coherent plan. Thus, making sense of an environment that includes other agents entails making sense of the other agents' actions in terms of what they are intended to achieve. When those actions are utterances, the utterances must be understood as actions in a plan the agents are trying to effect. That is, the speaker's plan must be recognized—the intentional account.

Generally, when a speaker says something it is with the goal of the hearer believing the content of the utterance, or thinking about it, or considering it, or taking some other cognitive stance toward it. Let us subsume all these mental terms under the term "cognize". Then we can summarize the relation between the intentional and informational accounts succinctly in the following formula:

$$\textbf{intentional-account} = goal(A, cognize(B, \textbf{informational-account})$$

The speaker ostensibly has the goal of changing the mental state of the hearer to include some mental stance toward the content characterized by the informational account. Thus, the informational account is embedded in the intentional account. When we reason about the speaker's intention, we are reasoning about how this goal fits into the larger picture of the speaker's ongoing plan. We are asking why the speaker seems to be trying to get the hearer to believe this particular content. The informational account explains the situation described in the discourse; the intentional account explains why the speaker chose to convey this information.

The (defeasible) axiom that encapsulates this is

$$\begin{aligned}&(\forall\, s,h,e_1,e,w)goal(s,e_1)\ \wedge\ cognize'(e_1,h,e)\ \wedge\ Segment(w,e)\\&\quad\supset\ utter(s,h,w)\end{aligned}$$

That is, normally if a speaker $s$ has a goal of the hearer $h$ cognizing a situation $e$ and $w$ is a string of words that conveys $e$, then $s$ will utter $w$ to $h$. We appeal to this axiom to interpret the utterance as an intentional communicative act. That is, if you ($U$) utter to me ($I$) a string of words ($W$), then to explain this observable event, I have to prove

$$utter(U,I,W)$$

and I begin to do so by backchaining on the above axiom. Reasoning about the speaker's plan is a matter of establishing the first two propositions in the antecedent of the axiom. Determining the informational content of the utterance is a matter of establishing the third, as described in the previous sections. The two sides of the proof influence each other since they share variables and since minimality results when both are explained and when they share propositions.

Both the intentional and informational accounts are necessary. The informational account is needed because we have no direct access to the speaker's plan. We can only infer it from history and behavior. The content of the utterance is often the best evidence of the speaker's intention, and often the intention is no more than to convey that particular content. On the other hand, the intentional account is necessary in cases like pragmatic ellipsis, where the informational account is highly underdetermined and the global interpretation is primarily shaped by our beliefs about the speaker's plan.

Perhaps most interesting are cases of genuine conflict between the two accounts. The informational account does not seem to be true, or it seems to run counter to the speaker's goals for the hearer to come to believe it, or it ought to be obvious that the hearer already does believe it. Tautologies are an example of the last of these cases—tautologies such as "boys will be boys," "fair is fair," and "a job is a job." Norvig and Wilensky (1990) cite this figure of speech as something that should cause trouble for an abduction approach that seeks minimal explanations, since the minimal explanation is that they just express a known truth. Such an explanation requires no assumptions at all.

In fact, the phenomenon is a good example of why an informational account of discourse interpretation has to be embedded in an intentional account. Let us imagine two parents, A and B, sitting in the playground and talking.

A: Your Johnny is certainly acting up today, isn't he?
B: Boys will be boys.

In order to avoid dealing with the complications of plurals and tense in this example, let us simplify B's utterance to

B: A boy is a boy.

Several informational accounts of this utterance are possible. The first is the Literal Extensional Interpretation. The first "a boy" introduces a specific, previously unidentified boy and the second says about him that he is a boy. The second informational account is the Literal Intensional Interpretation. The sentence expresses a trivial implicative relation between two general propositions—$boy(x)$ and $boy(x)$. The third is the Desired Interpretation. The first "a boy" identifies the typical member of a class which Johnny is a member of and the second conveys a general property, "being a boy", as a way of conveying a specific property, "misbehaving", which is true of members of that class.

More precisely, the logical form of the sentence can be written as follows:

$$(\exists\, e_1, e_2, x, y, z, w) boy'(e_1, x) \wedge rel(z, x) \wedge be(z, w) \wedge rel(w, y) \wedge \\ boy'(e_2, y)$$

The sentence expresses a *be* relation between two entities, but either or both of its arguments may be subject to coercion. Thus, we have introduced the two *rel* relations. The logical form can be given the tortured paraphrase, "$z$ is $w$, where $z$ is related to $x$ whose boy-ness is $e_1$ and $w$ is related to $y$ whose boy-ness is $e_2$."

The required axioms are as follows:

Everything is itself:

$$(\forall\, x) be(x,x)$$

Implication can be expressed by "to be":

$$(\forall\, e_1,e_2) imply(e_1,e_2) \supset be(e_1,e_2)$$

Implication is reflexive:

$$(\forall\, e) imply(e,e)$$

Boys misbehave:

$$(\forall\, e_1,x) boy'(e_1,x) \supset (\exists\, e_3) misbehave'(e_3,x) \wedge imply(e_1,e_3)$$

Misbehavers are often boys:

$$(\forall\, e_3,x) misbehave'(e_3,x) \wedge etc_1(x) \supset (\exists\, e_2) boy'(e_2,x)$$

Identity is a possible coercion relation:

$$(\forall\, x) rel(x,x)$$

An entity can be coerced into a property of the entity:

$$(\forall\, e,x) boy'(e,x) \supset rel(e,x)$$
$$(\forall\, e,x) misbehave'(e,x) \supset rel(e,x)$$

Note that we have axioms in both directions relating boys and misbehaving; in Section 3.3.10 the general way of expressing axioms is with biconditionals and *etc* predicates. The axioms with the coercion relation *rel* in the consequent begin to spell out the range of possible interpretations for *rel*.

Now the Literal Extensional Interpretation is established by taking the two coercion relations to be identity, taking *be* to be expressing identity, and assuming $boy(e_1,x)$ (or equivalently, $boy(e_2,y)$).

In the Literal Intensional Interpretation, $z$ is identified with $e_1$, $w$ is identified with $e_2$, and $boy'(e_1,x)$ and $boy'(e_2,y)$ are taken to be the two coercion relations. Then $e_2$ is identified with $e_1$ and $be(e_1,e_1)$ is interpreted as a consequence of $imply(e_1,e_1)$. Again, $boy(e_1,x)$ is assumed.

In the Desired Interpretation, the first coercion relation is taken to be $boy'(e_1,x)$, identifying $z$ as $e_1$. The second coercion relation is taken to be $misbehave'(e_3,y)$, identifying $w$ as $e_3$. If $etc_1(y)$ is assumed, then $misbehave'(e_3,y)$ explains $boy(e_2,y)$. If $boy(e_1,x)$ is assumed, it can explain $misbehave'(e_3,y)$, identifying $x$ and $y$, and also $imply(e_1,e_3)$. The latter explains $be(e_1,e_3)$.

** FIGURE **

Considering the informational account alone, the Literal Extensional Interpretation is minimal and hence would be favored. The Desired Interpretation is the worst of the three.

But the Literal Extensional and Intensional Interpretations leave the *fact* that the utterance *occurred* unaccounted for. In the intentional account, this is what we need to explain. The explanation would run something like this:

> B wants A to believe that B is not responsible for Johnny's misbehaving.
>
> Thus, B wants A to believe that Johnny misbehaves necessarily.
>
> Thus, given that Johnny is necessarily a boy, B wants A to believe that Johnny's being a boy implies that he misbehaves.
>
> Thus, B wants to convey to A that being a boy implies misbehaving.
>
> Thus, given that boy-ness implies misbehaving is a possible interpretation of a boy being a boy, B wants to say to A that a boy is a boy.

The content of the utterance under the Literal Extensional and Intensional Interpretations do not lend themselves to explanations for the fact that the utterance occurred, whereas the Desired Interpretation does. The requirement for the *globally* minimal explanation in an intentional account, that is, the requirement that both the content and the fact of the utterance must be explained, forces us into an interpretation of the content that would not be favored in an informational account alone. We are forced into an interpretation of the content that, while not optimal locally, contributes to a global interpretation that *is* optimal.

## 3.9 Weighted Abduction

In deduction, from $(\forall x)p(x) \supset q(x)$ and $p(A)$, one concludes $q(A)$. In induction, from $p(A)$ and $q(A)$, or more likely, from a number of instances of $p(A)$ and $q(A)$ together with other facts, one concludes $(\forall x)p(x) \supset q(x)$. Abduction is the third possibility. From $(\forall x)p(x) \supset q(x)$ and $q(A)$, one concludes $p(A)$. One can think of $q(A)$ as the observable evidence, of $(\forall x)p(x) \supset q(x)$ as a general principle that could explain $q(A)$'s occurrence, and of $p(A)$ as the inferred, underlying cause or explanation of $q(A)$. Of course, this mode of inference is not valid; there may be many possible such $p(A)$'s. Therefore, other criteria are needed to choose among the possibilities. We have already argued that abduction is the way to understand discourse; but because multiple inconsistant proofs and thus interpretations are possible, we need a way of choosing the best one.

One obvious criterion is the consistency of $p(A)$ with the rest of what one knows. Two other criteria are what Thagard (1978) has called *simplicity* and *consilience*. Roughly, simplicity is that $p(A)$ should be as small as possible, and consilience is that $q(A)$ should be as big as possible. We want to get more bang for the buck, where $q(A)$ is bang, and $p(A)$ is buck.

There is a property of natural language discourse, noticed by a number of linguists (e.g., Joos, 1972; Wilks, 1972), that suggests a role for simplicity and consilience in interpretation—its high degree of redundancy. Consider

> Inspection of oil filter revealed metal particles.

An inspection is a looking at that *causes one to learn* a property relevant to the *function* of the inspected object. The *function* of a filter is to capture *particles* from a fluid. To reveal is to *cause one to learn*. If we assume the two causings to learn are identical, the two sets of particles are identical, and the two functions are identical, then we have explained the sentence in a minimal fashion. Because we have exploited this redundancy, a small number of inferences and assumptions (simplicity) have explained a large number of syntactically independent propositions in the sentence (consilience). As a by-product, we have moreover shown that the inspector is the one to whom the particles are revealed and that the particles are in the filter, facts which are not explicitly conveyed by the sentence.

Another issue that arises in abduction in choosing among potential explanations is what might be called the "informativeness-correctness tradeoff". Many previous uses of abduction in AI from a theorem-proving perspective have been in diagnostic reasoning (e.g., Pople, 1973; Cox and Pietrzykowski, 1986), and they have assumed "most-specific abduction". If we wish to explain chest pains, it is not sufficient to assume the cause is simply chest pains. We want something more specific, such as "pneumonia". We want the most specific possible explanation. In natural language processing, however, we often want the least specific assumption. If there is a mention of a fluid, we do not necessarily want to assume it is lube oil. Assuming simply the existence of a fluid may be the best we can do.[1] However, if there is corroborating evidence, we may want to make a more specific assumption. In

[1] As Freud is purported to have said, "Sometimes a cigar is just a cigar."

Alarm sounded. Flow obstructed.

if we know the alarm is for the lube oil pressure, then this provides evidence that the flow is not merely of a fluid but of lube oil. The more specific our assumptions are, the more informative our interpretation is. The less specific they are, the more likely they are to be correct.

We therefore need a scheme of abductive inference with three features. First, it should be possible for goal expressions to be assumable, at varying costs. Second, there should be the possibility of making assumptions at various levels of specificity. Third, there should be a way of exploiting the natural redundancy of texts to yield more economic proofs.

Weighted abduction just such a scheme.[2] First, every conjunct in the logical form of the sentence is given an assumability cost. Second, this cost is passed back to the antecedents in Horn clauses by assigning weights to them. Axioms are stated in the form

$$(3.1) \qquad P_1^{w_1} \wedge P_2^{w_2} \supset Q$$

This says that $P_1$ and $P_2$ imply $Q$, but also that if the cost of assuming $Q$ is $c$, then the cost of assuming $P_1$ is $w_1 c$, and the cost of assuming $P_2$ is $w_2 c$.[3] Third, factoring or synthesis is allowed. That is, goal expressions may be unified, in which case the resulting expression is given the smaller of the costs of the input expressions. Thus, if the goal expression is of the form

$$(\exists \ldots, x, y, \ldots) \ldots \wedge q(x) \wedge \ldots \wedge q(y) \wedge \ldots$$

where $q(x)$ costs \$20 and $q(y)$ costs \$10, then factoring assumes $x$ and $y$ to be identical and yields an expression of the form

$$(\exists \ldots, x, \ldots) \ldots \wedge q(x) \wedge \ldots$$

where $q(x)$ costs \$10. This feature leads to minimality through the exploitation of redundancy.

Note that in (3.1), if $w_1 + w_2 < 1$, most-specific abduction is favored—why assume $Q$ when it is cheaper to assume $P_1$ and $P_2$. If $w_1 + w_2 > 1$, least-specific abduction is favored—why assume $P_1$ and $P_2$ when it is cheaper to assume $Q$. But in

[2] The abduction scheme is due to Mark Stickel, and it, or a variant of it, is described at greater length in Stickel (1988, 1989).

[3] Stickel (1989) generalizes the weights to arbitrary functions of $c$.

$$P_1^{.6} \wedge P_2^{.6} \supset Q$$

if $P_1$ has already been derived, it is cheaper to assume $P_2$ than $Q$. $P_1$ has provided evidence for $Q$, and assuming the "balance" $P_2$ of the necessary evidence for $Q$ should be cheaper.

Factoring can also override least-specific abduction. Suppose we have the axioms

$$P_1^{.6} \wedge P_2^{.6} \supset Q_1$$
$$P_2^{.6} \wedge P_3^{.6} \supset Q_2$$

and we wish to derive $Q_1 \wedge Q_2$, where each conjunct has an assumability cost of \$10. Assuming $Q_1 \wedge Q_2$ will then cost \$20, whereas assuming $P_1 \wedge P_2 \wedge P_3$ will cost only \$18, since the two instances of $P_2$ can be unified. Thus, the abduction scheme allows us to adopt the careful policy of favoring least-specific abduction while also allowing us to exploit the redundancy of texts for more specific and thus more informative interpretations.

Finally, we should note that whenever an assumption is made, it first must be checked for consistency, at least in a shallow manner.

In the above examples we have used equal weights on the conjuncts in the antecedents. It is more reasonable, however, to assign the weights according to the "semantic contribution" each conjunct makes to the consequent. Consider, for example, the axiom

$$(\forall\, x) car(x)^{.8} \wedge \textit{no-top}(x)^{.4} \supset \textit{convertible}(x)$$

We have an intuitive sense that *car* contributes more to *convertible* than *no-top* does. We are more likely to assume something is a convertible if we know that it is a car than if we know it has no top.[4] The weights on the conjuncts in the antecedent are adjusted accordingly.

Exactly how the weights and costs should be assigned and the sematics of the number are explored in Chapter 7. Until then it will be shown how the rules allow correct interpretations to be *possible*, and occasionally it will be argued that the correct interpretation is *likely* to be chosen because of redundancy in the text. But arguments that the correct interpretation will be the one *selected* from among all the possible interpretations will have to await Chapter 7.

[4] To prime this intuition, imagine two doors. Behind one is a car. Behind the other is something with no top. You pick a door. If there's a convertible behind it, you get to keep it. Which door would you pick?

## 3.10 "Et Cetera" Propositions and the Form of Axioms

In the abductive approach to interpretation, we determine what implies the logical form of the sentence rather than determining what can be inferred from it. We backward-chain rather than forward-chain. Thus, one would think that we could not use superset information in processing the sentence. Since we are backward-chaining from the propositions in the logical form, the fact that, say, lube oil is a fluid, which would be expressed as

$$(\forall\, x)\mathit{lube\text{-}oil}(x) \supset \mathit{fluid}(x) \tag{3.2}$$

could not play a role in the analysis of a sentence containing "lube oil". This is inconvenient. In the text

> Flow obstructed. Metal particles in lube oil filter.

we know from the first sentence that there is a fluid. We would like to identify it with the lube oil mentioned in the second sentence. In interpreting the second sentence, we must prove the expression

$$(\exists\, x)\mathit{lube\text{-}oil}(x)$$

If we had as an axiom

$$(\forall\, x)\mathit{fluid}(x) \supset \mathit{lube\text{-}oil}(x)$$

then we could establish the identity. But of course we don't have such an axiom, for it isn't true. There are lots of other kinds of fluids. There would seem to be no way to use superset information in our scheme.

Fortunately, however, there is a way. We can make use of this information by converting the axiom to a biconditional. In general, axioms of the form

$$\text{species} \supset \text{genus}$$

can be converted into biconditional axioms of the form

$$\text{genus} \wedge \text{differentiae} \equiv \text{species}$$

Often as in the above example, we will not be able to prove the differentiae, and in many cases the differentiae cannot even be spelled out. But in our abductive scheme, this does not matter; they can simply be assumed. In fact, we need not state them explicitly. We can simply introduce a predicate which stands for all the remaining properties. It will never be provable, but it will be assumable. Thus, we can rewrite (3.2) as

$$(3.3) \qquad (\forall\, x) fluid(x)^{.6} \;\wedge\; etc_1(x)^{.6} \;\equiv\; lube\text{-}oil(x)$$

Then the fact that something is fluid can be used as evidence for its being lube oil, since we can assume $etc_1(x)$. With the weights distributed according to semantic contribution, we can go to extremes and use an axiom like

$$(\forall\, x) mammal(x)^{.2} \;\wedge\; etc_2(x)^{.9} \;\supset\; elephant(x)$$

to allow us to use the fact that something is a mammal as (weak) evidence for its being an elephant. This axiom can be taken to say, "One way of being a mammal is being an elephant."

Although this device may seem ad hoc, we view it as implementing a fairly general solution to the problems of nonmonotonicity in commonsense reasoning and vagueness of meaning in natural language. The use of "et cetera" propositions is a very powerful, and liberating, device. Before we hit upon this device, in our attempts at axiomatizing a domain in a way that would accommodate many texts, we were always "arrow hacking"—trying to figure out which way the implication had to go if we were to get the right interpretations, and lamenting when that made no semantic sense. With "et cetera" predications, that problem went away, and for principled reasons. Implicative relations could be used in either direction. Moreover, their use is liberating when constructing axioms for a knowledge base. It is well-known that almost no concept can be defined precisely. We are now able to come as close to a definition as we can and introduce an "et cetera" proposition with an appropriate weight to indicate how far short we feel we have fallen.

The "et cetera" propositions play a role analogous to the abnormality propositions of circumscriptive logic (McCarthy, 1980, 1987). In circumscriptive theories it is usual to write axioms like

$$(\forall\, x) bird(x) \;\wedge\; \neg Ab_1(x) \;\supset\; flies(x)$$

This certainly looks like the axiom

$$(\forall\, x) bird(x) \;\wedge\; etc_3(x)^{w} \;\supset\; flies(x)$$

The literal $\neg Ab_1(x)$ says that $x$ is not abnormal in some particular respect. The literal $etc_3(x)$ says that $x$ possesses certain unspecified properties, for example, that $x$ is not abnormal in that same respect. In circumscription, one minimizes over the abnormality predicates, assuming they are false wherever possible, perhaps with a partial ordering on abnormality predicates to determine which assumptions to select (e.g., Poole, 1989). Our abduction scheme generalizes this a bit: The literal $etc_3(x)$ may be assumed if no contradiction results and if the resulting proof is the most economical one available. Moreover, the "et cetera" predicates can be used for any kind of differentiae distinguishing a species from the rest of a genus, and not just for those related to normality.

There is no particular difficulty in specifying a semantics for the "et cetera" predicates. Formally, $etc_1$ in axiom (3.3) can be taken to denote the set of all things that either are not fluid or are lube oil. Intuitively, $etc_1$ conveys all the information one would need to know beyond fluidness to conclude that something is lube oil. As with nearly every predicate in an axiomatization of commonsense knowledge, it is hopeless to spell out necessary and sufficient conditions for an "et cetera" predicate. In fact, the use of such predicates is motivated largely by a recognition of this fact about commonsense knowledge.

The "et cetera" predicates could be used as the abnormality predicates are in circumscriptive logic, with separate axioms spelling out conditions under which they would hold. However, in the view adopted here, more detailed conditions would be spelled out by expanding axioms of the form

$$(\forall\, x)p_1(x) \;\wedge\; etc_4(x) \;\supset\; q(x)$$

to axioms of the form

$$(\forall\, x)p_1(x) \;\wedge\; p_2(x) \;\wedge\; etc_5(x) \;\supset\; q(x)$$

where the weight on $etc_5(x)$ would be less than that on $etc_4(x)$. An "et cetera" predicate would appear only in the antecedent of a single axiom and never in a consequent. Thus, the "et cetera" predications are only place-holders for assumption costs. They are never proved. They are only assumed.

Let us summarize at this point the most elaborate form axioms in the knowledge base will have. If we wish to express an implicative relation between concepts $p$ and $q$, the most natural way to do so is as the axiom

$$(\forall\, x, z)p(x, z) \;\supset\; (\exists\, y)q(x, y)$$

where $z$ and $y$ stand for arguments that occur in one predication but not in the other. When we introduce eventualities, this axiom becomes

$$(\forall\, e_1, x, z) p'(e_1, x, z) \;\supset\; (\exists\, e_2, y) q'(e_2, x, y)$$

Next we introduce an "et cetera" proposition into the antecedent to take care of the imprecision of our knowledge of the implicative relation.

$$(\forall\, e_1, x, z) p'(e_1, x, z) \;\wedge\; etc_1(x, z) \;\supset\; (\exists\, e_2, y) q'(e_2, x, y)$$

It is also useful to make explicit in the axiom itself the implication relation between the antecedent and the consequent. We can do this by including an *imply* predication in the consequent.

$$(\forall\, e_1, x, z) p'(e_1, x, z) \wedge etc_1(x, z) \;\supset\; (\exists\, e_2, y) q'(e_2, x, y) \wedge imply(e_1, e_2)$$

This relation can be made stronger in some cases. For example there may be a causal or enabling relation between the eventualities described in the antecedent and consequent. In these cases, the predicates *cause* or *enable* would replace *imply*. We saw this in the axioms of Section 3.3.7.

Since the rules are only defeasible, and especially since we are using *etc* predications to allow inference from a general term to a specific term, the relation *imply* is sometimes too strong. For example, being a mammal does not *imply* being an elephant. Thus, we will sometimes also want to *weaken* the *imply* relation to mere association—the predicate *assoc*.

Now we can biconditionalize the relation between $p$ and $q$ by writing the converse axiom as well:

$$\begin{aligned}
&(\forall\, e_1, x, z) p'(e_1, x, z) \wedge etc_1(x, z) \;\supset\; (\exists\, e_2, y) q'(e_2, x, y) \wedge imply(e_1, e_2)\\
&(\forall\, e_1, x, y) q'(e_2, x, y) \wedge etc_2(x, y) \;\supset\; (\exists\, e_1, z) p'(e_1, x, z) \wedge assoc(e_2, e_1)
\end{aligned}$$

This then is the most general formal expression in our abductive logic of what is intuitively felt to be an *association* between the concepts $p$ and $q$.

In this book, for notational convenience, I will use the simplest form of axiom I can get away with for the example. The reader should keep in mind however that these are only abbreviations for the full, biconditionalized form of the axiom.[5]

[5] The full axioms are non-Horn, but not seriously so. They can be Skolemized and broken into two axioms having the same Skolem functions. This remark holds as well for other axioms in this article that have conjunctions in the consequent.

## 3.11 Anchoring Abduction in a Structured Connectionist Model

Because of its elegance and very broad coverage, the "Interpretation as Abduction" model is very appealing on the symbolic level. But to be a plausible candidate for how people understand language, there must be an account of how it could be implemented in neurons. In this section I describe in outline how the abduction framework can be realized in a structured connectionist model called SHRUTI developed by Lokendra Shastri (Ajjanagadde and Shastri, 1991; Shastri and Ajjanagadde, 1993; Shastri and Grannes, 1996; Shastri, 1999; Shastri and Wendelken, 2000; Wendelken and Shastri, 2002). Of course, substantial work still remains to determine whether this kind of model is what actually exists in the human brain, although there is suggestive evidence. But by linking the symbolic and connectionist levels we are at least providing a proof of *possibility* for the "Interpretation as Abduction" framework.[6]

Traditional connectionist models have been very good at implementing defeasible *propositional* logic. Indeed, all the applications to natural language processing in this tradition begin by setting up the problem so that it is a problem in propositional logic. But this is not adequate for natural language understanding in general. For example, the coreference problem requires the expressivity of first-order logic to even state. We need a way of expressing predicate-argument relations and a way of expressing different instantiations of the same general principle. We need a mechanism for universal instantiation. In the connectionist literature, this has gone under the name of the *variable-binding problem.*

Prior to the development of SHRUTI it was considered difficult (McCarthy, 1988) and even impossible (Fodor and Pylyshyn, 1988) for connectionist models to solve in a biologically plausible manner the complex version of the variable-binding problem arising in the systematic propagation of dynamic bindings for instantiating inferred relations. These misgivings were understandable because conventional techniques for representing variable bindings require storing and communicating names or pointers, but the storage and processing capacity of neurally plausible nodes and the resolution of their outputs are insufficient to support this functionality. SHRUTI offers a neurally plausible solution to the variable binding problem based on temporal synchrony.

The essential idea behind the SHRUTI architecture is simple and elegant.

[6] Parts of this section were written in collaboration with Lokendra Shastri.

A predication is represented as an assemblage or cluster of nodes, and axioms representing general knowledge are realized as connections among these clusters. Inference is accomplished by means of spreading activation through these structures.

In the cluster representing predications, two nodes, a collector node and an enabler node, correspond to the predicate and fire asynchronously. The level of activation on the enabler node keeps track of the "utility" of this predication in the proof that is being searched for. That is, the activation is higher the greater the need to find a proof for this predication. The level of activation on the collector node is higher the greater the plausibility that this predication is part of the desired proof. We can think of the activations on the enabler nodes as prioritizing goal expressions, whereas the activations on the collector nodes indicate degree of belief in the predications, or more properly, degree of belief in the current relevance of the predications. The connections between nodes of different predication clusters have a strength of activation, or link weight, that corresponds to strength of association between the two concepts. The proof process then consists of activation spreading through enabler nodes, as we backchain through axioms, and then spreading back through collector nodes after bottoming out in something known.

In addition, in the predication cluster, there are argument nodes, one for each argument of the predication. These fire synchronously with the argument nodes in other predication clusters to which they are connected. Thus, if the clusters for $p(x,y)$ and $q(z,x)$ are connected, with the two $x$ nodes linked to each other, then the two $x$ nodes will fire in synchrony, and the $y$ and $z$ nodes will fire at an offset with the $x$ nodes and with each other. This synchronous firing indicates that the two $x$ nodes represent variables bound to the same value. This constitutes the solution to the variable-binding problem.

Proofs are searched for in parallel, and winner-takes-all circuitry suppresses all but the one whose collector nodes have the highest level of activation.

There are complications in this model for such things as managing different predications with the same predicate but different arguments. But the essential idea is as described. In brief, the view of relational information processing implied by SHRUTI is one where reasoning is a transient but systematic propagation of *rhythmic* activity over structured cell-ensembles, each active entity is a phase in the rhythmic activity, dynamic bindings are represented by the *synchronous* firing of appropriate nodes, long-term facts are circuits that detect coincidences in the ongoing flux of rhythmic activity, and rules are high-efficacy links that cause the propagation of rhythmic activity between cell-ensembles. Reasoning is the spontaneous outcome of a SHRUTI network.

In the abduction framework, the typical axiom in the knowledge base is of the form

$$(\forall\, x, y)[p_1(x, y) \wedge p_2(x, y) \supset (\exists\, z)[q_1(x, z) \wedge q_2(x, z)]] \tag{3.4}$$

That is, the top level logical connective will be implication. There may be multiple predications in the antecedent and in the consequent. There may be variables ($x$) that occur in both the antecedent and the consequent, variables ($y$) that occur only in the antecedent, and variables ($z$) that occur only in the consequent. Abduction backchains from predications in consequents of axioms to predications in antecedents. Every step in the search for a proof can be considered an abductive proof where all unproved predications are assumed for a cost. The best proof is the least cost proof.

The implementation of this axiom in SHRUTI requires predication clusters of nodes, as described in Section 2, and axiom clusters of nodes (see Figure 3.6). A predication cluster has one collector node and one enabler node, both firing asynchronously, corresponding to the predicate and one synchronously firing node for each argument. An axiom cluster has one collector node and one enabler node, both firing asynchronously, recording the plausibility and the utility, respectively, of this axiom participating in the best proof. It also has one synchronously firing node for each variable in the axiom – in our example, nodes for $x$, $y$ and $z$.

The axiom is then encoded in a structure like that shown in Figure 2. There is a predication cluster for each of the predications in the axiom and one axiom cluster that links the predications of the consequent and antecedent. In general, the predication clusters will occur in many axioms; this is why their linkage in a particular axiom must be mediated by an axiom cluster.

Suppose the proof process is backchaining from the predication $q_1(x, z)$. The activation on the enabler node (?) of the cluster for $q_1(x, z)$ induces an activation on the enabler node for the axiom cluster. This in turn induces activation on the predication nodes for $p_1(x, y)$ and $p_2(x, y)$. Meanwhile the firing of the $x$ node in the $q_1$ cluster induces the $x$ node of the axiom cluster to fire in synchrony with it, which in turn causes the $x$ nodes of the $p_1$ and $p_2$ clusters to fire in synchrony as well. In addition, a link (not shown) from the enabler node of the axiom cluster to the $y$ argument node of the same

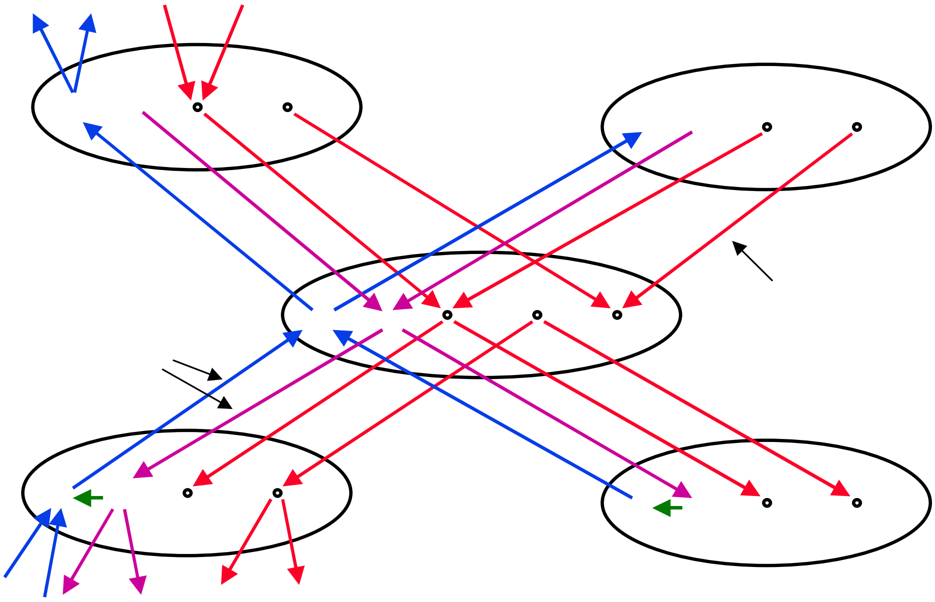

Figure 3.6: Axiom encoded in SHRUTI.

cluster causes the $y$ argument node to fire, while links (not shown) from the $x$ and $z$ nodes cause that firing to be out of phase with the firing of the $x$ and $z$ nodes. This firing of the $y$ node of the axiom cluster induces synchronous firing in the $y$ nodes of the $p_1$ and $p_2$ clusters.

By this means we have backchained over axiom (9) while keeping distinct the variables that are bound to different values. We are then ready to backchain over axioms in which $p_1$ and $p_2$ are in the consequent.

As mentioned above, the $q_1$ cluster is linked to other axioms as well, and in the course of backchaining, it induces activation in those axioms' clusters too. In this way, the search for a proof proceeds in parallel. Inhibitory links will eventually force a winner-takes-all outcome.

The strength of activation from the $q_1$ enabler node to the axiom enabler node is proportional to the conditional probability that this axiom is used in proving $q_1(x, z)$.

A link from the enabler to the collector node of a predication cluster transmits activation proportional to the prior probability of an instance of this predication being true and relevant and inversely proportional to the strength of activation on the enabler node. This is how proofs bottom out, either in assumptions or in predications known to hold (high prior). The inverse relation between enablers and collectors is because a high activation on enablers indicates a high utility of proving rather than assuming the predication.

This description conveys the essence of how abductive inference can be implemented in SHRUTI, although it is oversimplified in several respects. Certain technical problems arise, and how they are dealt with is addressed in Chapter 7.

## 3.12 Learning Structured Knowledge

### 3.12.1 Incremental Changes to Axioms

To have a plausible biological model of language understanding, we must also have a biologically plausible account of how the required knowledge can be learned. Here we present such an account, first at the symbolic level in terms of incremental changes to axioms and then at the connectionist level in terms of node recruitment.

Most work in learning has assumed that the basic structure of the knowledge base is known and the problem is to adjust weights to maximize the frequency of the right answers. The current implementation of SHRUTI already supports this form of learning. The harder problem in learning is to learn the structure of the axioms. Our view is that axioms can be built up in an incremental fashion, as long as this involves a biologically plausible mechanism and each increment confers an advantage. Axioms can be constructed incrementally through sequences of five symbolic operations that each have neural realizations in the SHRUTI framework using the mechanism of *recruitment* learning (see below). The five operations are as follows:

1. Increase the arity of a predicate: $p(x) \Rightarrow p(x,y)$

   For example, it may be learned that *mother* is not a property of a single entity but a relation between two entities: $mother(x) \Rightarrow mother(x,y)$

2. Introduce a new predicate $p_1$ as a specialization of an old predicate $p$: $p_1(x) \supset p(x)$

   For example, the predicate *beagle* may be introduced as a specialization of *dog*: $beagle(x) \supset dog(x)$

3. Add a proposition to the antecedent of an axiom:
   $p_1(x) \supset q(x) \Rightarrow p_1(x) \wedge p_2(x) \supset q(x)$

For example, it may be learned that not every seat is a chair, but only seats with backs:
$seat(x) \supset chair(x) \Rightarrow seat(x) \wedge back(y,x) \supset chair(x)$

4. Add a proposition to the consequent of an axiom:
$p(x) \supset q_1(x) \Rightarrow p(x) \supset q_1(x) \wedge q_2(x)$

   For example, it may be learned that snow is not only white, but also cold.
$snow(x) \supset white(x) \Rightarrow snow(x) \supset white(x) \wedge cold(x)$

5. Telescope or chunk two axioms to produce a shorter inferential path (as in SOAR (Newell, 1990)):
$p_1(x) \supset p_2(x); p_2(x) \supset p_3(x) \Rightarrow p_1(x) \supset p_3(x)$

Two other incremental changes can be viewed as special cases of 3 and 4 above.

- Specialize predicates in the antecedent of an axiom:
$p(x) \supset q(x) \Rightarrow p_1(x) \supset q(x),$
where $p_1(x) \supset p(x)$

  For example, it may be learned that not all dogs have floppy ears, only beagles. $dog(x) \supset \textit{floppy-ears}(x) \Rightarrow beagle(x) \supset \textit{floppy-ears}(x),$

- Specialize predicates in the consequent of an axiom:

  $p(x) \supset q(x) \Rightarrow p(x) \supset q_1(x),$
where $q_1(x) \supset q(x)$

  For example, it may be learned that rabbits are not only mammals; more specifically, they are rodents. $rabbit(x) \supset mammal(x) \Rightarrow rabbit(x) \supset rodent(x),$

Each of these incremental operations clearly refines the knowledge base and hence confers greater functionality. We now discuss how it could be implemented at the neural level.

### 3.12.2 Recruitment Learning

[7]

[7] This section was largely written by Lokendra Shastri.

Recruitment learning (Wickelgren, 1979; Feldman, 1982; Shastri, 1988; Die-derich, 1989; Valiant, 1994) can be described as follows: Learning occurs within a network of randomly connected nodes. Recruited nodes are those nodes in the network that have acquired a distinct meaning or functionality by virtue of their strong interconnections to other recruited nodes and/or other sensorimotor (i.e., input/output) nodes. Nodes that are not yet recruited can be viewed as “free” nodes. Such nodes are connected via weak links to a large number of free, recruited, and/or sensorimotor nodes. These free nodes form a primordial network from which suitably connected nodes may be recruited for representing new items. For example, a novel concept $q$ that can be expressed as a conjunct of existing concepts $p_1$ and $p_2$ can be learned by (i) identifying free nodes that receive links from nodes representing $p_1$ as well as nodes representing $p_2$ and (ii) “recruiting” one or more such free nodes by strengthening the weights of links incident on such nodes from $p_1$ and $p_2$ nodes.

The notion of recruitment learning has been extended to include the formation of relational concepts (Shastri, 1988; Valiant, 1994) and it has been shown that recruitment learning can be firmly grounded in the biological phenomena of *long-term potentiation* (LTP) and *long-term depression* (LTD) that involve rapid, long-lasting, and highly specific changes in synaptic strength (Shastri, 2001).

Wendelken and Shastri (2000) show how the + to + links and the ? to ? links with appropriate probabilistic weights can be formed among relational clusters to encode *causal* knowledge, and a similar method can be used for implementing the five incremental changes to axioms.

A central feature of causation is that causes precede effects. It is therefore reasonable to assume that we would only learn a causal rule $A \rightarrow B$ when we observe A to occur before B. As a learning rule for SHRUTI, this can mean that a +A to+B link should be strengthened only when the activity at the source of the link (+A) precedes activity at the target (+B). For a ? to ? link, we assume the opposite behavior, namely that such a link should be strengthened if and only if its source becomes active once its target is already firing. In this manner, we require that one observation precede another by some amount of time and also that both occur within a specified window. Recently, it has been shown that this form of learning which depends on the relative timing of spikes in pre- and post-synaptic cells is biologically plausible and has interesting computational properties (see Song, Miller and Abbot, 2000).

The starting point is a collection of relational focal clusters, differentiated into + nodes, ? nodes, and argument nodes. It is assumed that each type of node is connected via low weight links to a large number of nodes of its own type; this connection would typically be mediated by an intermediate layer of cells. The learning rule for a + to + link is as follows: if the source has been active sufficiently long and the target then becomes active, the link weight is updated as $w_{t+1} = w_t + \alpha * (1 - w_t)$ where $\alpha = 1/\#updates$; otherwise if the source has been active sufficiently long and the target fails to fire, the weight is decremented $w_{t+1} = w_t - \alpha * w_t$. If the target becomes active before the source, then there is no change. It is easy to see that $w_{t+1} = (\#increases)/(\#updates)$, where $w_0 = 0$, correctly encodes the probability that the target of the link follows the source within a specified window of time. A modification of this rule including a normalization term, to account for the possibility of multiple sources, reduces the weight increase on a link by a factor proportional to the number of active links impinging on the same target (Grossberg, 1987). This allows a link weight for $+A \rightarrow +B$ to encode $P(B|onlyA)$ and not just $P(B|A)$.

For ? to ? links, the learning rule is nearly the reverse of the above. In this case, a similar weight increase occurs whenever a link target has been active for sufficiently long and a source becomes active, and a weight decrease occurs when a target remains active for too long without activity at the source. If the source becomes active first, then there is no change. Again, it is easy to see that this link correctly records the probability that the source fires after the target (within a designated time window), and for a link $?B \rightarrow ?A$ this can be reasonably interpreted as $P(B|A)$.

Links between arguments, the links connecting the ? and + nodes of a predicate, and inhibitory links also have to be learned in a suitable manner.

Accounts similar to this can be constructed for each of the incremental operations. For example, adding a new consequent to an axiom involves recruiting a new predication cluster and strengthening its connections to the collector, enabler and argument nodes of the axiom cluster, in response to frequent simultaneous activation of the two clusters. These issues are discussed further in Chapter 7.

## 3.13 Relation to Relevance Theory

One of the other principal contenders for a theory of how we understand extended discourse is Relevance Theory (RT) (Sperber and Wilson, 1986). In fact, the "Interpretation as Abduction" (IA) framework and RT are very close to each other in the processing that would implement them.

In RT, the agent is in the situation of having a knowlege base $K$ and hearing a sentence with content $Q$. From $K$ and $Q$ a new set $R$ of inferences can be drawn:

$$K, Q \vdash R$$

RT says that the agent strives to *maximize* $R$ in an appropriately hedged sense. An immediate consequence of this is that insofar as we are able to pragmatically strengthen $Q$ by means of axioms of the form

$$P \supset Q$$

then we are getting a better $R$, since $P$ implies anything that $Q$ implies, and then some. In the IA framework, we begin with pragmatic strengthening. The task of the agent is to explain the general $Q$ with the more specific $P$.

This means that anything done in the IA framework ought to carry over without change into RT. Much of the work in RT depends primarily or solely on pragmatic strengthening, and where this is the case, it can immediately be incorporated into the IA framework. RT also makes use of pragmatic loosening. In Chapter 6 it is shown how the same effect can be achieved via the coercion used in interpreting metonymy.

From the point of view of IA, people are going through the world trying to figure out what is going on. From the point of view of RT, they are going through the world trying to learn as much as they can, and figuring out what is going on is in service of that.

The IA framework has been worked out in greater detail formally and, I believe, has a more compelling justification—explaining the observables in our environment. But a great deal of excellent work has been done in RT, so it is useful to know that the two frameworks are almost entirely compatible.

## 3.14 The Golden Spike

Because of its complexity, the problem of understanding how the brain works is immense. It is simply not possible to describe directly how neurobiological mechanisms result in intelligent behavior. A common strategy in cognitive science is to divide the problem into subproblems by hypothesizing two intermediate levels – a symbolic level and a connectionist level. Intelligent behavior is implemented in a symbolic level, which is in turn implemented in a connectionist architecture, which is realized in assemblages of neurons.

There has been substantial success in artificial intelligence and other fields in implementing varieties of intelligent behavior in symbolic architectures. In the last decade and a half, a consensus has been emerging that abduction, or inference to the best explanation, is a key process underlying much intelligent behavior (e.g., Josephson and Josephson, 1990). In this book I show that a very broad range of problems in natural language understanding can be viewed as abductive inference. To interpret a text one must find a minimal proof of the logical form of the sentences in the text, allowing for assumptions, and linguistic problems such as the resolution of reference, metonymy, and syntactic ambiguity are solved as a by-product of such a process. More generally, with the right logical formalizations of discourse structure, the rules of syntax, and commonsense knowledge, the interpretation of discourse can be viewed as a matter of coming up with the "best" proof that the string of words constitutes a coherent block of discourse, allowing assumptions, where "best" is determined by a cost function. Heretofore the problems with symbolic models have been that they lack biological plausibility, there are no compelling accounts of how learning could take place within them, and the means of choosing among alternatives, such as the cost function in abduction, tend to be ad hoc.

The appeal of connectionist architectures is precisely that they solve these three problems. The operation of nodes in connectionist models is biologically motivated; it is based on what we know about the way neurons communicate. There is a natural story to tell about how learning takes place in connectionist models. Connectionist models are designed to choose among alternatives. Typically, however, in order for them to exhibit anything approaching intelligent behavior, they have to be set up in specialized ways for the problem being solved. They have generally lacked the expressive power required for natural language understanding. While they are able to choose among competing hypotheses well, there has been no natural way for them to do, for example, coreference resolution and other tasks involving reasoning about the identity and nonidentity of entities. Essentially, they implement a kind of "soft" propositional logic, whereas intelligent behavior requires a kind of "soft" first-order logic. There needs to be a way of expressing general rules and keeping different instances of a general rule distinct, which in first-order logic is realized by universal instantiation.

Shastri has developed the connectionist model SHRUTI that solves this problem dynamically by means of the synchronous firing of neurons. That is, logical axioms are represented in assemblages of nodes, with different nodes representing predicates and arguments. When argument nodes are bound to the same value, they fire in synchrony. While it is impossible to know in the current state of neurobiology whether this model is correct, we do know that synchronous firing of neurons occurs and that it has something to do with entities being in the same category. The model is thus at least biologically plausible.

This book explicates how the the abductive model of natural language understanding can be linked up with the SHRUTI connectionist model. SHRUTI provides the biological plausibility that symbolic models lack. A natural method for the automatic learning of axioms is suggested. SHRUTI provides a way of choosing among alternative interpretations discovered by abductive reasoning. At the same time, the abductive framework gives the connectionist model the expressive power required for realistically complex intelligent behavior.

In the enterprise of linking intelligent behavior with neurobiology via symbolic and connectionist levels, the connection between the two intermediate levels has heretofore been missing. By linking them here, this work points the way toward that Golden Spike that will connect accounts of intelligent actions with accounts of biological activity.

## Acknowledgements

This material is based in part on work supported by the National Science Foundation and Advanced Research Projects Agency under Grant Number IRI-9304961 (Integrated Techniques for Generation and Interpretation), and by the National Science Foundation under Grant Number IRI-9619126 (Multimodal Access to Spatial Data). Any opinions, findings, and conclusions or recommendations expressed in this chapter are those of the author and do not necessarily reflect the views of the National Science Foundation.

# Bibliography

[1] Appelt, Douglas E., and Martha E. Pollack, 1990. "Weighted Abduction for Plan Ascription", Technical Note 491, SRI International, Menlo Park, California, May 1990.

[2] Charniak, Eugene, 1986. "A Neat Theory of Marker Passing", *Proceedings, AAAI-86, Fifth National Conference on Artificial Intelligence*, Philadelphia, Pennsylvania, pp. 584-588.

[3] Charniak, Eugene, and Robert Goldman, 1988. "A Logic for Semantic Interpretation", *Proceedings, 26th Annual Meeting of the Association for Computational Linguistics*, pp. 87-94, Buffalo, New York, June 1988.

[4] Charniak, Eugene, and Robert Goldman, 1989. "A Semantics for Probabilistic Quantifier-Free First-Order Languages, with Particular Application to Story Understanding", *Proceedings, Eleventh International Joint Conference on Artificial Intelligence*, pp. 1074-1079. Detroit, Michigan. August 1989.

[5] Charniak, Eugene, and Drew McDermott, 1985. *Introduction to Artificial Intelligence.* Reading, Mass.: Addison-Wesley Publishing Co.

[6] Cox, P. T., and T. Pietrzykowski, 1986. "Causes for Events: Their Computation and Applications", *Proceedings*, Eighth International Conference on Automated Deduction(CADE-8), pp. 608-621, Oxford, England.

[7] Dasigi, Venu R., 1988. *Word Sense Disambiguation in Descriptive Text Interpretation: A Dual-Route Parsimonious Covering Model* (doctoral dissertation), Technical Report TR-2151, Department of Computer Science, University of Maryland, College Park, December, 1988. Also published as Technical Report WSU-CS-90-03, Department of Computer Science and Engineering, Wright State University, Dayton, Ohio.

[8] Diederich, Joaquim, 1989. “Instruction and High-Level Learning in Connectionist Networks”. *Connection Science*, Vol. 1, pp. 161-180.

[9] Feldman, Jerome A., 1982. “Dynamic Connections in Neural Networks, *Bio-Cybernetics*, Vol. 46, pp. 27-39.

[10] Fodor, Jerry A., and Zenon W. Pylyshyn, 1988. “Connectionism and Cognitive Architecture: A Critical Analysis”. In S. Pinker and J. Mehler, Eds., *Connections and Symbols*, MIT Press.

[11] Ginsberg, Matthew L., editor, 1987. *Readings in Nonmonotonic Reasoning*, Morgan Kaufmann Publishers, Inc., Los Altos, California.

[12] Grice, H. P., 1967. “Logic and Conversation”, William James Lectures, Harvard University, manuscript.

[13] Grice, H. P., 1989. “Logic and Conversation”, in *Studies in the Ways of Words*, Harvard University Press, Cambridge, MA, pp. 22-40.

[14] Grossberg, Stephen, 1987. “Competitive Learning: From Interactive Adaptation to Adaptive Resonance”. *Cognitive Science*, Vol. 11, pp. 23-26.

[15] Harabagiu, Sanda, and Dan Moldovan, 1998. “Knowledge Processing on an Extended WordNet”, in C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, pp. 379-405. MIT Press, 1998.

[16] Harabagiu, S. and D.I. Moldovan. 2002. “LCC’s Question Answering System”. *Proceedings*, 11th Text Retrieval Conference, TREC-11. Gaithersburg, MD.

[17] Hewitt, Carl E., 1972. “Description and Theoretical Analysis (using schemas) of PLANNER: A Language for Proving Theorems and Manipulating Models in a Robot”, Technical Report TR-258, AI Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

[18] Hirst, Graeme, 1987. *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press, Cambridge, England.

[19] Hobbs, Jerry R., Mark Stickel, Paul Martin, and Douglas Edwards, 1988. “Interpretation as Abduction”, *Proceedings, 26th Annual Meeting of the Association for Computational Linguistics*, pp. 95-103, Buffalo, New York, June 1988.

[20] Hobbs, Jerry R., Mark Stickel, Douglas Appelt, and Paul Martin, 1993. "Interpretation as Abduction", *Artificial Intelligence*, Vol. 63, Nos. 1-2, pp. 69-142.

[21] Joos, Martin, 1972. "Semantic Axiom Number One", *Language*, pp. 257-265.

[22] Josephson, John R. and Susan G. Josephson, 1990. *Abductive Inference: Computation, Philosophy, Technology*, Cambridge, England.

[23] Lakatos, Imre, 1970. "Falsification and the Methodology of Scientific Research Programmes", in *Criticism and the Growth of Knowledge*, ed. Imre Lakatos and Alan Musgrave. Cambridge University Press, Cambridge, England, pp. 91-196.

[24] Lascarides, Alex, and Jon Oberlander, 1992. "Abducing Temporal Discourse", in R. Dale, E. Hovy, D. Rosner, and O. Stock, eds., *Aspects of Automated Natural Language Generation*, Springer, Berlin, pp. 167–182.

[25] Lewis, David, 1979. "Scorekeeping in a Language Game," *Journal of Philosophical Logic*, Vol. 6, pp. 339-59.

[26] Mann, William, and Sandra Thompson, 1986, "Relational Propositions in Discourse", *Discourse Processes*, Vol. 9, No. 1, pp. 57-90.

[27] McCarthy, John, 1980. "Circumscription: A Form of Nonmonotonic Reasoning", *Artificial Intelligence*, Vol. 13, pp. 27-39. Reprinted in M. Ginsberg, ed., *Readings in Nonmonotonic Reasoning*, pp. 145-152, Morgan Kaufmann Publishers, Inc., Los Altos, California.

[28] McCarthy, John, 1987. "Circumscription: A Form of Nonmonotonic Reasoning", in M. Ginsberg, ed., *Readings in Nonmonotonic Reasoning*, pp. 145-152, Morgan Kaufmann Publishers, Inc., Los Altos, California.

[29] McCarthy, John, 1988. "Epistemological Challenges for Connectionism. Commentary to 'Proper Treatment of Connectionism' by Paul Smolensky", *Behavioral and Brain Sciences*, Vol. 11, No. 1.

[30] McDermott, Drew, and John Doyle, 1980. "Non-monotonic Logic I", *Artificial Intelligence*, Vol. 13, Nos. 1,2. pp. 41-72. April 1980.

[31] McRoy, Susan, and Graeme Hirst, 1991. "An Abductive Account of Repair in Conversation", *Working Notes*, AAAI Fall Symposium on Discourse Structure in Natural Language Understanding and Generation, Asilomar, California, November 1991, pp. 52-57.

[32] Morgan, Charles G., 1971. “Hypothesis Generation by Machine”, *Artificial Intelligence*, Vol. 2, pp. 179-187.

[33] Nagao, Katashi, 1989. “Semantic Interpretation Based on the Multi-World Model”, In *Proceedings of the Eleventh International Conference on Artificial Intelligence*, pp. 1467-1473, Detroit, Michigan.

[34] Newell, Allen, 1990. *Unified Theories of Cognition.* Harvard University Press.

[35] Newton, Isaac, 1934 [1686]. *Mathematical Principles of Natural Philosophy*, Vol. 1: *The Motion of Bodies*, and Vol. 2: *The System of the World*, translated by Andrew Motte and Florian Cajori, University of California Press, Berkeley, California.

[36] Ng, Hwee Tou and Raymond J. Mooney, 1990. “The Role of Coherence in Abductive Explanation”. *Proceedings*, Eighth National Conference on Artificial Intelligence, pp. 337-342, Boston, MA, August 1990.

[37] Norvig, Peter, 1983. “Frame Activated Inferences in a Story Understanding Program”, *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, pp. 624-626.

[38] Norvig, Peter, 1987. “Inference in Text Understanding”, *Proceedings, AAAI-87, Sixth National Conference on Artificial Intelligence*, pp. 561-565, Seattle, Washington, July 1987.

[39] Norvig, Peter, and Robert Wilensky, 1990. “A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation”, in H. Karlgren, ed., *Proceedings*, Thirteenth International Conference on Computational Linguistics, Helsinki, Finland, Vol. 3, pp. 225-230, August, 1990.

[40] Pearl, Judea, 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., San Mateo, California.

[41] Pierce, Charles Sanders, 1955. “Abduction and Induction”, in Justus Buchler, editor, *Philosophical Writings of Pierce*, pp. 150-156, Dover Books, New York.

[42] Pollard, Carl, and Ivan A. Sag, 1994. *Head-Driven Phrase Structure Grammar*, University of Chicago Press and CSLI Publications.

[43] Pople, Harry E., Jr., 1973, "On the Mechanization of Abductive Logic", *Proceedings, Third International Joint Conference on Artificial Intelligence*, pp. 147-152, Stanford, California, August 1973.

[44] Rayner, Manny, 1993. *Abductive Equivalential Translation and its application to Natural Language Database Interfacing*, Ph.D. thesis, Royal Institute of Technology, Stockholm, September 1993.

[45] Reggia, James A., 1985. "Abductive Inference", in K. N. Karna, editor, *Proceedings of the Expert Systems in Government Symposium*, pp. 484-489, IEEE Computer Society Press, New York.

[46] Reggia, James A., Dana S. Nau, and Pearl Y. Wang, 1983. "Diagnostic Expert Systems Based on a Set Covering Model", *International Journal of Man-Machine Studies*, Vol. 19, pp. 437-460.

[47] Reiter, Raymond, and Giovanni Criscuolo, 1981. "On Interacting Defaults", *Proceedings*, Seventh International Joint Conference on Artificial Intelligence, pp. 270-276, Vancouver, BC, August 1981.

[48] Shastri, Lokendra, 1988. "A Connectionist Approach to Knowledge Representation and Limited Inference", *Cognitive Science*, Vol. 12, No. 3, pp. 331-392.

[49] Shastri, Lokendra, 2001. "A Computational Model of Episodic Memory Formation in the Hippocampal System", *Neurocomputing*, Vol. 38-40, pp. 889-897. Also appeared in Bower, J. Ed., *Computational Neuroscience, Trends in Research 2001*, Elsevier, Amsterdam.

[50] Shastri, Lokendra, and Venkat Ajjanagadde, 1993. "From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables and Dynamic Bindings Using Temporal Synchrony", *Behavioral and Brain Sciences*, Vol. 16, pp. 417-494.

[51] Shastri, Lokendra, and Carter Wendelken, 2000. "Seeking Coherent Explanations — A Fusion of Structured Connectionism, Temporal Synchrony, and Evidential Reasoning", *Proceedings*, Twenty Second Annual Conference of the Cognitive Science Society, pp. 453-458, Philadelphia, PA. August 2000. Available at http://www.icsi.berkeley.edu/ shastri/ShrutiSmritiRefs.html

[52] Shoham, Yoav, 1987. "Nonmonotonic Logics: Meaning and Utility", *Proceedings, International Joint Conference on Artificial Intelligence*, pp. 388-393. Milano, Italy, August 1987.

[53] Song, Sen, Kenneth D. Miller, and L. F. Abbot, 2000. "Competitive Hebbian Learning Through Spike-Timing Dependent Synaptic Plasticity". *Nature Neurosciece*, Vol. 3, pp. 919-926.

[54] Sperber, Dan, and Deirdre Wilson, 1986. *Relevance: Communication and Cognition*, Harvard University Press, Cambridge, Massachusetts.

[55] Stickel, Mark E., 1988. "A Prolog-like Inference System for Computing Minimum-Cost Abductive Explanations in Natural-Language Interpretation", *Proceedings of the International Computer Science Conference-88*, pp. 343–350, Hong Kong, December 1988. Also published as Technical Note 451, Artificial Intelligence Center, SRI International, Menlo Park, California, September 1988.

[56] Stickel, Mark E., 1989. "Rationale and Methods for Abductive Reasoning in Natural-Language Interpretation", in R. Studer (ed.). *Proceedings*, Natural Language and Logic, International Scientific Symposium, Hamburg, Germany, May 1989, *Lecture Notes in Artificial Intelligence* #259, Springer-Verlag, Berlin, pp. 233–252.

[57] Thagard, Paul R., 1978. "The Best Explanation: Criteria for Theory Choice", *The Journal of Philosophy*, pp. 76-92.

[58] Thomason, Richmond H., 1990. "Accommodation, Meaning, and Implicature: Interdisciplinary Foundations for Pragmatics", in *Intentions in Communication*, P. Cohen, J. Morgan, and M. Pollack, editors, Bradford Books (MIT Press), Cambridge, Massachusetts, pp. 325-364.

[59] Valiant, Leslie G., 1994. *Circuits of the Mind*, Oxford University Press, New York.

[60] Wendelken, Carter, and Lokendra Shastri, 2000. "Probabilistic Inference and Learning in a Connectionist Causal Network". *Proceedings*, Neural Computation 2000, Berlin, May 2000.

[61] Wendelken, Carter, and Lokendra Shastri, 2002. "Combining Belief and Utility in a Structured Connectionist Agent Architecture". *Proceedings*, Twenty Fourth Annual Conference of the Cognitive Science Society, pp. 926-931. Available at http://www.icsi.berkeley.edu/ shastri/ShrutiSmritiRefs.html

[62] Wickelgren, Wayne A., 1979. "Chunking and Consolidation: A Theoretical Synthesis of Semantic Networks, Configuring in Conditioning, S-R

Versus Cognitive Learning, Normal Forgetting, the Amnesic Syndrome, and the Hippocampal Arousal System". *Psychological Review*, Vol. 86, No. 1, pp. 44-60.

[63] Wilensky, Robert, 1983. *Planning and Understanding: A Computational Approach to Human Reasoning*, Addison-Wesley, Reading, Massachusetts.

[64] Wilensky, Robert, David N. Chin, Marc Luria, James Martin, James Mayfield, and Dekai Wu, 1988. "The Berkeley UNIX Consultant Project", *Computational Linguistics*, vol. 14, no. 4, December 1988, pp. 35-84.

[65] Wilks, Yorick, 1972. *Grammar, Meaning, and the Machine Analysis of Language*, Routledge and Kegan Paul, London.

[66] Winograd, Terry, 1972. *Understanding Natural Language*, Academic Press, New York.

[67] Wolff, Christian, 1963 [1728]. *Preliminary Discourse on Philosophy in General*, R. J. Blackwell (trans.), Bobbs-Merrill, Indianapolis IN.