2 Knowledge of Meaning and Theories of Truth

We have been pursuing semantics as a theory of the real but unconscious knowledge of speakers. We have argued that what is known by speakers is a set of rules and principles that are finite in number and compositional in form. These underlie our grasp of semantic facts, our capacity to make semantic judgments, and our ability to communicate with and understand others. Precisely what does this knowledge consist in? What kinds of rules and principles are known?

The idea we will adopt and develop in this book derives from the work of Donald Davidson, who proposes that the work of a semantic theory can be done by a particular sort of formal theory called a **truth theory**, or **T theory** for short.[1] A T theory for a particular language L is a deductive system that has the resources to prove something about the truth value of every sentence of L. More specifically, for each sentence of L it proves a theorem of the form in (T), where S is a name or description of the L sentence and p has the same truth value as the sentence referred to by S:

(T) S is true if and only if p.

The language from which S is drawn is typically referred to as the **object language**. It is the language that we are theorizing about. Here the object language is L. The language used to discuss the object language is typically referred to as the **metalanguage**. It is the language in which our theory is stated. Here the metalanguage is English. A T theory produces theorems that pair a sentence S of the object language L with a sentence p of the metalanguage. These two are paired together by the relation "is true if and only if" (henceforth abbreviated "is true iff").[2] Theorems of the form (T) are called **T sentences** or **T theorems**.[3]

2.1 T Theories

The easiest way to understand the workings of a T theory is to examine a concrete instance. So let us consider a sample T theory for a small sublanguage of English that we will call PC, since it includes some elements of the propositional calculus. PC contains an infinite number of sentences, although they are all of a highly restricted form. In particular, PC contains the three elementary sentences *Phil ponders*, *Jill knows Kate*, and *Chris agrees*. PC also contains all sentences that can be produced from the basic three either by joining them together with one of the sentence conjunctions *and* or *or*, or by prefixing them with the negation *it is not the case that*. For present purposes, we will assume that the elementary sentences are generated by the three rules in (1), and that the remainder are generated by the rules in (2). Rule (1a) may be read as saying that *Phil ponders* is a sentence (an S), and similarly for (1b, c). Rule (2a) may be read as stating that any two sentences (Ss) joined by the word *and* is also a sentence, and similarly for (2b, c):[4]

(1) a. S → *Phil ponders*
    b. S → *Chris agrees*
    c. S → *Jill knows Kate*

(2) a. S → S *and* S
    b. S → S *or* S
    c. S → *It is not the case that* S

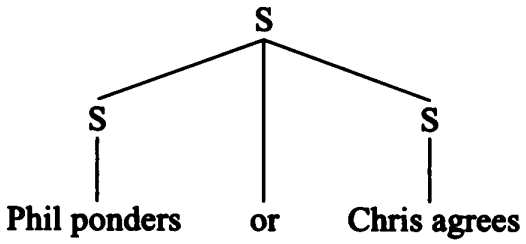
Under these rules PC will contain all of the example sentences in (3) (among infinitely many others):

(3) a. [$_S$ Phil ponders]
    b. [$_S$ Chris agrees]
    c. [$_S$ Jill knows Kate]
    d. [$_S$ [$_S$ Phil ponders] or [$_S$ Chris agrees]]
    e. [$_S$ [$_S$ Jill knows *Kate*] and [$_S$ [$_S$ Phil ponders] or [$_S$ Phil ponders]]]
    f. [$_S$ It is not the case that [$_S$ Jill knows Kate]]
    g. [$_S$ It is not the case that [$_S$ [$_S$ Phil ponders] or [$_S$ Chris agrees]]]

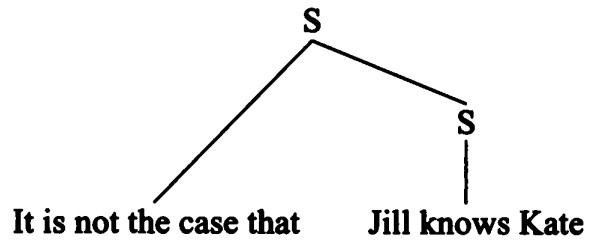
The **labeled brackets** in (3) depict the derivational histories of complex sentences. For example, (3g) is formed by our first connecting *Phil ponders* and *Chris agrees*, using *or* to form a disjunction (as permitted by (2b)) and then attaching *it is not the case that* to form the negation of this disjunction (as

permitted by (2c)). These derivations may be also depicted by means of familiar tree diagrams or **phrase markers.**[5] Thus (3d) can be associated with the tree in (4a). Likewise, (3f) can be represented with the tree in (4b):

(4) a.



b.



The three elementary sentences *Phil ponders*, *Jill knows Kate*, and *Chris agrees* appearing as the leaves on these trees are treated as complex words in PC. That is, they are assigned no internal syntactic structure. Following standard terminology, we will refer to these elementary sentences as the **terminal nodes** in the tree. And we will refer to nodes with internal syntactic structure, for example, the built-up Ss in (3d–g), as **nonterminal nodes.**

A T theory for PC will allow us to derive a T theorem for each sentence of PC. The T theory we wish to explore consists of three basic parts. First, there are interpretation rules for the terminal nodes in the PC grammar. These assign a semantic contribution to the basic components of PC, as in (5):

(5) a. *Phil ponders* is true iff Phil ponders.
b. *Chris agrees* is true iff Chris agrees.
c. *Jill knows Kate* is true iff Jill knows Kate.

Second, there are interpretation rules for the nonterminal nodes. These allow us to derive T sentences for configurations with internal structure from the T sentences for their smaller component sentences. Thus for *any* sentences S, S_1, and S_2,[6]

(6) a. $[_S\ S_1$ *and* $S_2]$ is true iff both S_1 is true and S_2 is true
b. $[_S\ S_1$ *or* $S_2]$ is true iff either S_1 is true or S_2 is true
c. $[_S$ *It is not the case that* S] is true iff it is not the case that S is true
d. $[_S\ \alpha]$ is true iff α is true (for any elementary sentence α)

Third and finally, there are **production rules**, which license inferences of certain specified kinds. These allow us to reason from the elementary and general semantic rules and to prove results using them. For PC we will adopt two production rules. The first is called substitution of equivalents, abbreviated (SE). This rule is defined as follows:

(SE) Substitution of equivalents

$$\frac{F(\alpha) \quad \alpha \text{ iff } \beta}{F(\beta)}$$

According to (SE), if we have proved a statement involving α (i.e., $F(\alpha)$) and have proved that α is equivalent to β, (i.e., α iff β), then we may conclude the result of substituting β for α in the statement (that is, we may conclude $F(\beta)$). The second rule is universal instantiation, or (UI):

(UI) Universal instantiation

$$\frac{\text{For any S, } F(\text{S})}{F(\alpha)}$$

Universal instantiation will allow us to apply the general rules in (6) to particular instances.

The three elements specified above may be viewed as jointly specifying a deductive system. The interpretation rules in (5) and (6) function as semantic axioms from which we can prove a T sentence for each sentence of PC, using the production rules (SE) and (UI).

2.1.1 A Sample Derivation

We can illustrate how this system works with example (3d). Recall this string of words receives the structure in (4a), which we represent with labeled bracketing as in (7):

(7) $[_S$ $[_S$ Phil ponders] or $[_S$ Chris agrees]]

To derive the T sentence for this structure, we begin at the top (leftmost) S node and interpret it in terms of its subsentences. We do this by applying (UI) to the axiom in (6b):

(8) $[_S$ $[_S$ Phil ponders] or $[_S$ Chris agrees]] is true iff either
$[_S$ Phil ponders] is true or $[_S$ Chris agrees] is true (by (6b), (UI))

The axiom in (6d) tells us how further to unpack each of the two disjuncts on the right-hand side of "iff" in (8):

(9) a. [$_S$ Phil ponders] is true iff *Phil ponders* is true (by (6d), (UI))
b. [$_S$ Chris agrees] is true iff *Chris agrees* is true (by (6d), (UI))

The clauses in (5a) and (5c) then tell us how to spell out the T sentence for each of the two elementary sentences on the right-hand sides of the "iff" connectors in (9):

(10) a. *Phil ponders* is true iff Phil ponders. (by (5a))
b. *Chris agrees* is true iff Chris agrees. (by (5b))

To complete the T sentence for (3d), we now use the production rule (SE), working our way back up the tree. Suppose that we let (9a) be $F(\alpha)$, "*Phil ponders* is true" be α, and "Phil ponders" be β. By (10a), we have "α iff β," so we can use (SE) to conclude "[$_S$ Phil ponders] is true iff Phil ponders":

(11) [$_S$ Phil ponders] is true iff *Phil ponders* is true (9a)
Phil ponders is true iff Phil ponders (10a)

[$_S$ Phil ponders] is true iff Phil ponders
(by (9a), (10a), (SE))

We now reapply the same strategy, this time letting (9b) be $F(\alpha)$, letting "*Chris agrees* is true" be α, and letting "Chris agrees" be β. By (10b), we have that "α iff β," so again, using (SE), we may conclude that (12b):

(12) a. [$_S$ Phil ponders] is true iff Phil ponders (by (9a), (10a), (SE))
b. [$_S$ Chris agrees] is true iff Chris agrees (by (9b), (10b), (SE))

Next let (8) be $F(\alpha)$, let "[$_S$ Phil ponders] is true" be α, and let "Phil ponders" be β. By (12a), we have "α iff β," so we can use (SE) to conclude that (13):

(13) [$_S$ [$_S$ Phil ponders] or [$_S$ Chris agrees]] is true iff either
Phil ponders or [$_S$ Chris agrees] is true (by (8), (12a), (SE))

Finally, let (13) be $F(\alpha)$, let "[$_S$ Chris agrees] is true" be α, and let "Chris agrees" be β. By (12b), we have "α iff β," so, using (SE), we conclude that (14):

(14) [$_S$ [$_S$ Phil ponders] or [$_S$ Chris agrees]] is true iff either
Phil ponders or Chris agrees (by (13), (12b), (SE))

We have shown that the PC sentence *Phil ponders or Chris agrees* is true if and only if Phil ponders or Chris agrees. Intuitively, of course, this is a true outcome.

2.1.2 The Nontriviality of T Theorems

Results like that in (14) are achieved by means of a formally precise theory using explicit procedures. Nonetheless, they may appear entirely trivial at first sight. After all, how informative is it to learn that *Phil ponders or Chris agrees* is true if and only if Phil ponders or Chris agrees? In fact, such results are quite informative and the initial impression of triviality is really an illusion of sorts. It is important to locate the source of this illusion and to dispel it.

Like other scientific statements, T theorems state facts or hypotheses about certain phenomena, specifically, about linguistic phenomena. T theorems thus make statements *in* a language *about* a language: language is mentioned in the T theorem, and it is used to formulate the T theorem. Given any object-language, there will be a vast array of languages in which we can formulate T sentences for sentences of that object language. We can give the truth conditions of a sentence of English using a T sentence formulated in Chinese. We can also give the truth conditions of a sentence of English using a T sentence formulated in English. This is the case with the T theory for PC. The object language, PC, is a small fragment of English. And the metalanguage in which we give our axioms, production rules, and results is also English.

Now it is a simple fact about sentences and their truth conditions that any sentence of a language L can be used to state its own truth conditions in L. That is to say, T sentences of the following homophonic form are (almost) always true for any sentence S of English:[7]

(T*) *S* is true if and only if S.

It is partly this fact that gives results like (14) their air of triviality. The T sentence (14) mentions a sentence of English (*Phil ponders or Chris agrees*) and then goes on to use this very sentence to state its truth conditions. Part of what is obvious about T theorems like (14) is thus their truth. Nonetheless, it is important to see that despite the obvious truth of (14), this T sentence is far from trivial.

The general informativeness of truth theories can be seen most clearly when we consider examples where the object language and metalanguage diverge. For example, consider an alternative T theory for PC in which the meta-

language is German rather than English. Under this choice, (5) and (6) would be replaced by (5′) and (6′):

(5′) a. *Phil ponders* ist wahr genau dann wenn Phil nachdenkt.
b. *Chris agrees* ist wahr genau dann wenn Chris zustimmt.
c. *Jill knows Kate* ist wahr genau dann wenn Jill Kate kennt.

(6′) a. [$_S$ S_1 *and* S_2] ist wahr genau dann wenn S_1 wahr ist und S_2 wahr ist.
b. [$_S$ S_1 *or* S_2] ist wahr genau dann wenn S_1 ist wahr oder S_2 wahr ist.
c. [$_S$ *It is not the case that* S] ist wahr genau dann wenn S nicht wahr ist.
d. [$_S$ α] ist wahr genau dann wenn S wahr ist (fur jeden Elementarsatz α)

These results are evidently neither uninformative nor trivial. For a monolingual speaker of German, (5′) and (6′) will provide information about an infinite set of English sentences. Using these rules, the German speaker will be able to determine the truth conditions for all of the sentences comprising PC. However, (5′) and (6′) do not say anything different from what is said by the original (5) and (6). Indeed, they say just the same thing! They attribute the same truth conditions to the same elementary sentences of PC, and they provide the same rules for dealing with complex sentences of PC. It's just that they say these things in German, and so make them available to monolingual German speakers.

Why, then, do homophonic T sentences appear trivial? The answer is not too hard to find. If someone is in a position to understand a T sentence, then, obviously, they must understand the metalanguage, the language in which the T sentence is formulated. Or, more precisely, they must understand as much of the metalanguage as is used in the T sentence. For example, to understand any T sentence of the form of (14), it is necessary to understand as much English as is used in its formulation:

(15) S is true iff Phil ponders.

Understanding this includes understanding the RHS, "Phil ponders." Now anyone who understands the sentence *Phil ponders* knows that it is true if and only if Phil ponders. Understanding *Phil ponders* requires this at the very least. Consequently, anyone who is in a position to understand any T sentence of the form of (15) already knows what is stated by (5a): *Phil ponders* is true if and only if Phil ponders. But, of course, (5a) just is a T sentence of the form of (15). So anyone who understands (5a) must already know that it is true.

In a sense, then, homophonic T sentences are uninformative: anyone who is in a position to understand them already knows enough to know that they

Table 2.1 Truth table for the material biconditional, "iff"

p	*q*	*p* iff *q*
t	t	t
t	f	f
f	t	f
f	f	t

are true. But it certainly does not follow from this that what is stated by such a T sentence is not highly substantive. It is highly substantive, as we can see by the nonhomophonic cases. The purportedly unsubstantive (5a) says no more and no less than the evidently substantive (5'a).[8]

Although T sentences are not trivial, at the same time we should emphasize that by themselves they carry less information than one might think. In a T sentence the "if and only if" that appears is just the ordinary **material biconditional**, defined in standard logic texts by the truth table in table 2.1. According to this table, a sentence made up of two sentences *p* and *q* joined by "iff" is true whenever *p* and *q* are either both true or both false (and false otherwise). Since the "is true" in a T sentence is just the usual predicate "true," any T sentence in which the sentence on the left of "iff" has the same truth value as the RHS will be true. For example, (16) is perfectly true.

(16) *Snow is white* is true iff pigs have curly tails.

The upshot is that although T sentences carry nontrivial semantic information relating the truth value of a sentence to a worldly condition, the information they carry is somewhat limited. As we will shortly see, the issue of exactly how much information a T sentence carries is a central one.

2.2 T Theories as Theories of Meaning

We began our investigations by characterizing semantics as the study of linguistic meaning, more precisely, as the study of knowledge of meaning. From this perspective, it may seem quite puzzling to be told that knowledge of meaning amounts to knowledge of a T theory. After all, a T theory like the one for PC proves statements about the truth of sentences, for example, those

in (17). It proves nothing directly about the meanings of sentences; it does not give results like (18):

(17) a. *Phil ponders* is true iff Phil ponders.
b. *Phil ponders or Chris agrees* is true iff either Phil ponders or Chris agrees.

(18) a. *Phil ponders* means that Phil ponders.
b. *Phil ponders or Chris agrees* means that either Phil ponders or Chris agrees.

But then what is the relation between the two? How are we to understand the claim that knowledge of the one might be responsible for our grasp of the other?

As a step toward clarifying the connection, note first that although our T theory for PC doesn't derive explicit statements of meaning, its results are similar to the latter in an important respect. Observe that while the sentences in (17) differ from those in (18) in involving the relation "is true iff" instead of the relation "means that," the two sets of statements are alike in pairing exactly the same object language sentences and metalanguage sentences together. That is, the relation "is true iff" defined by our truth theory for PC (that is, by (5), (6), (SE), and (UI)) is similar to the relation "means that" insofar as it associates an object-language sentence with a metalanguage sentence that intuitively gives its meaning. We will call a T theory that yields the same pairing as that given by "means that" an **interpretive T theory**. And we will call the T sentences yielded by an interpretive T theory **interpretive T sentences**.

We propose that knowledge of this special kind of T theory, an interpretive T theory, is what underlies our grasp of semantic facts, our ability to understand our language in the way that we do. That is, we propose the following empirical hypothesis about knowledge of meaning:

The T hypothesis A speaker's knowledge of meaning for a language *L* is knowledge of a deductive system (i.e., a system of axioms and production rules) proving theorems of the form of (T) that are interpretive for sentences of *L*.

On this view, speakers who know the semantics of their language have internalized a system of rules like those in PC. The deliverances of this system, its (interpretive) T sentences, are what the speaker draws upon to encode and decode utterances, to make semantic judgments, and so on.

To explain the connection between interpretive T theories and semantic knowledge more fully and to show how the former could serve as a theory of meaning, it is useful to consider the T hypothesis in the light of two important questions that arise naturally in connection with it. First, under the T hypothesis, speakers are claimed to have internalized an interpretive T theory. But is it really possible to define such a T theory formally? Since "is true iff" and "means that" are different relations, is it possible to give a set of axioms and deductive principles for a natural language whose T theorems pair all and only those paired by "means that"? We call this the **extension question**.

Second, under the T hypothesis, knowledge of interpretive T theorems is claimed to provide the information that underwrites judgments about semantic facts. But would knowing an interpretive T theory be enough to tell you what the sentences of a language mean, to ground judgments about semantic properties and relations, and to account for the external and internal significance of language? Again, since "is true iff" and "means that" are two different relations that appear to talk about two very different things, truth versus meaning, a positive answer is by no means clear. We call this the **information question.**[9]

2.2.1 The Extension Question

There are two separate parts to the extension question. First, there is the question of whether we can give a T theory that's sufficiently productive, that is, one that proves an interpretive T theorem for every sentence of the object language. Second, there is the question of whether we can give a T theory that's not overproductive, that is, one that proves no uninterpretive results.

Both of these questions appear to be fundamentally empirical ones, to be answered by providing an interpretive T theory of the kind required. In subsequent chapters we will explore such theories for a wide variety of natural-language constructions, including such central ones as predicates, names, quantifiers, descriptions, anaphoric elements, modifiers, and embedded clauses. And there are other constructions not discussed here, such as comparatives, that have also been more or less satisfactorily dealt with. There remain, of course, elements of natural language that have so far resisted satisfactory T-theoretic treatment. Subjunctive conditionals are one well-known example. As we will see, however, the track record of T theories is a strong one—strong enough to warrant optimism about their power in principle to account for the full range of natural-language structures.

The second part of the extension question—whether it's possible to give a T theory that proves no uninterpretive results—is a bit more complex. As described earlier, a T theory consists of two basic parts: semantic axioms for interpreting lexical items and phrasal structures, and production rules for deducing results from the axioms. T theorems are the product of these components acting together, and including either the wrong axioms or the wrong production rules can yield uninterpretive results. For example, suppose that axiom (5c) (repeated below) were replaced with (19), or suppose that we simply added (19) to PC as it now stands:

(5c) *Jill knows Kate* is true iff Jill knows Kate.

(19) *Jill knows Kate* is true iff Jill knows Kate and 2 plus 2 equals 4.

Then we could obviously deduce uninterpretive T-theorems (including (19) itself). This is the case even though (19) is perfectly true, as are all the new T theorems that would result from its addition to the original theory.

A similar result holds with the production rules in the T theory for PC. This theory includes the highly restricted rule of substitution (SE) (repeated below). However, suppose that we replaced (SE) with the alternative rule (SE′).

(SE) Substitution of equivalents

$$\frac{F(\alpha) \quad \alpha \text{ iff } \beta}{F(\beta)}$$

(SE′) Substitution of equivalents, version 2

For any formula β such that α iff β,

$$\frac{F(\alpha)}{F(\beta)}$$

The two differ as follows: under (SE), we are allowed to substitute a formula β only if we have *proved* that it is equivalent to α as part of the derivation whereas under (SE′), we are allowed to substitute *any* β that is materially equivalent to α.[10] Such a change will once again yield uninterpretive results. For example, it is a fact of logic that the equivalence in (20) holds.

(20) Jill knows Kate iff Jill knows Kate and 2 plus 2 equals 4.

Accordingly, from this fact and interpretive axiom (5c), (SE′) will allow us to prove an uninterpretive T theorem, as shown in (21):

(21) *Jill knows Kate* is true iff Jill knows Kate. (5c)

Jill knows Kate is true
iff Jill knows Kate and 2 plus 2 equals 4. (by (20), (21a), (SE'))

The original production rule (SE) blocks this result because it does not allow substitution of arbitrary equivalents but only βs that have been proven equivalent to α as part of the derivation. This feature essentially encapsulates the derivation, blocking the importation of extraneous results like (20). Hence uninterpretive consequences like (21) are not derivable.

The contrast between (SE) and (SE') illustrates a further important point. Readers familiar with logic will note that the alternative substitution rule (SE') is not a bizarre one in any sense but is just the rule for substitution of material equivalents standardly assumed in logic texts.[11] What we see, then, is that one of the standard inference rules of logic is not admissible in our semantic theory. This result is in fact quite general. If one were to add to our T theory for PC even the ordinary, truth-preserving rules of standard logical systems, one would easily be able to prove uninterpretive consequences, such as the T theorem in (22):[12]

(22) *Chris agrees and Jill knows Kate* is true iff it is not the case that either it is not the case that Chris agrees or it is not the case that Jill knows Kate.

That the inference rules of logic allow too wide a class of deductions for semantic theory is not an accidental fact but rather follows from the very different goals of logic and semantics, as conceived here. As Frege (1956 [1918]) succinctly put it, logic is the general theory of truth. The inference rules of logic are motivated by the conceptual or philosophical goal of characterizing the set of true inferences from a given set of axioms. By contrast, semantics is a theory of speaker knowledge. The production rules of semantics are motivated by the empirical goal of describing part of what speakers know about a language. On the assumption that what speakers know about meaning is an interpretive T theory, it is clear that semantic production rules cannot aim to yield T sentences that are merely true. Rather, they must yield T sentences that are both true and interpretive, since these are what underlie knowledge of meaning, according to the T hypothesis.[13]

Since most familiar logical systems are not adequate for the purposes of semantics, it is an interesting but, at present, largely unexplored question as to what rules or production procedures should be employed in their place. The semantic theory given for PC restricts the class of deductions from its (inter-

pretive) lexical axioms to those derived by (UI) and (SE). As we will see in subsequent chapters, if the semantic axioms for the PC fragment are modified in even relatively simple ways, it becomes necessary to introduce additional production rules to prove the most basic interpretive results, and this immediately brings the risk of overgeneration. The production rules proposed in succeeding chapters do in fact produce only interpretive consequences. But there is no general theory at present of what formal procedures are best suited to the job of building interpretive T theories. This is an area of research that may ultimately draw on the resources of logic (proof theory) and psychology (reasoning and cognition).[14]

2.2.2 The Information Question

The T hypothesis not only assumes that it is possible to give an interpretive T theory for a language. It also asserts that knowledge of such a theory underlies judgments of actual meanings—that the information necessary for the latter is present in the former. This claim is strong and controversial. And, initially at least, it looks very dubious. The problem is that there just doesn't seem to be enough information in a T theory—even an interpretive one—to support judgments of meaning.

To see why this is so, consider the hypothetical situation in which you know no French but have at hand a T theory for French written out in English. Suppose that this theory is interpretive but that you don't know it is. You are now faced with the problem of finding out what certain French sentences mean, for example, the sentence in (23):

(23) Les oiseaux ont des plumes.

You can use the T theory to derive T theorems, such as (24):

(24) *Les oiseaux ont des plumes* is true iff birds have feathers.

But do you now know what any French sentence means? Do you know what *Les oiseaux ont des plumes* means? No. For (24) neither says nor implies (25), nor does anything else in the T theory.

(25) *Les oiseaux ont des plumes* means that birds have feathers.

The point here is a simple one. Since there are many true T theories for French in English that are not interpretive, you have no way of knowing in advance whether the particular theory before you is interpretive or not. And

since you don't know that the theory is interpretive, you can't move from (24) to (25). What goes here for explicit theories would seem to go for tacit ones as well. If you had tacit knowledge of an interpretive T theory for French, then you would not seem to know enough to know what French sentences mean.

It is tempting to think that this problem might be solved by specifying some new additional theory that, if known, would allow speakers to discover whether their T theory were interpretive or not. If you had a second, independent theory permitting you to deduce that the T theory yielding (24) was interpretive, then you could indeed pass from (24) to (25). On reflection, however, this does not seem to be a promising strategy. After all, a theory allowing you to deduce that (24) is interpretive would be one telling you that "birds have feathers" gives the meaning of *Les oiseaux ont des plumes*. But this is just the job we want our T theory to do! We wanted the T theory to correlate sentences with their meanings. Hence it looks like we could succeed in this strategy only at the price of putting our T theory out of a job.

T Theorems Treated as Interpretive

To motivate a more promising line on how a T theory might underlie judgments of meaning, let us recast the problem situation. In the previous little scenario, you had an interpretive T theory, but you didn't know that it was such. Because you didn't know this and couldn't deduce it, you didn't know you could use the theory to interpret French. Suppose, however, that because of your basic makeup you were compelled to take the T theory in your possession as interpretive, regardless of its actual status. Suppose, for example, that you are the kind of person who can't stand uncertainty, and to avoid it when dealing with Frenchmen, you simply decide to treat any T theory for French that comes into your hands as giving the meanings of French sentences. Whenever you are presented with a French sentence, you turn to your T theory, calculate a T sentence for it, and take whatever results on the RHS of the biconditional as giving its meaning.

Notice that in this circumstance, the knowledge gap noted earlier is still present: you still do not *know* that your T theory is an interpretive one, and the theory neither contains this information nor allows you to deduce it. Nonetheless, the T theory does underwrite judgments of meaning for you, since you use it to interpret and produce French sentences and you behave toward French speakers as if it rendered the meaning of their words. Given your constitution, you proceed as if the gap in question did not exist, as if you already knew your theory were interpretive. Notice furthermore that if the T

theory coming into your hands were in fact interpretive, then all your judgments and acts would be appropriate: you would correctly render the meanings of French sentences, you would form correct beliefs about what French speakers were trying to say to you, you would plan actions correctly in accord with these beliefs, and so on. Thus a T theory could serve as a theory of meaning for you, and do so successfully, if you were constituted to treat any T theory as interpretive and if events conspired to bring an interpretive T theory into your hands.

We suggest this second scenario as our answer to the information question —as a picture of how T theories might successfully serve as semantic theories for human speakers despite containing no explicit information about the meanings of expressions. Suppose that as a matter of biological endowment (that is, of universal grammar), humans are designed to acquire T theories. These theories are not written out in some natural language in a book but rather are represented internally in the brain. In the course of learning a language, speakers fix axioms and production rules yielding T theorems as outputs. Suppose further that humans are designed to treat whatever T-theory they acquire as interpretive. That is, whenever they are called upon to produce or interpret their own sentences or the sentences of others, they draw on the results of their internalized T theory and treat its theorems as interpretive: they take the RHS to give the meaning of the natural-language sentence mentioned on the LHS. Finally, suppose that events conspire to give speakers an interpretive T theory in the course of development. Then, although the T-theory contains no explicit information about the meanings of words and sentences, it would still be responsible for the semantic abilities of speakers. They would use it to make the judgments they do. The knowledge gap, though present, would be irrelevant to understanding or action, since speakers would proceed as if they already knew that their T theory were interpretive. Furthermore, since speakers would have learned an interpretive T theory as a matter of fact, all interpretations, beliefs, and actions undertaken in accordance with it would be appropriate ones. The use of the T theory as a theory of meaning would be successful.

The success of this proposal evidently turns on the correctness of its three central assumptions:

- Humans are designed to acquire a T theory.
- Humans are designed to treat any T theory they acquire as interpretive.
- In the course of development humans learn a T theory that is interpretive in fact.

About the first assumption we will say nothing more except that it is an obvious one, given our general approach. People acquire knowledge of meaning, and since we are assuming that a T theory underlies this knowledge, obviously we must assume that people acquire a T theory. As usual, checking the truth of such an assumption will be a highly indirect matter, involving the success of the larger enterprise as a whole.

The second assumption—that the mind treats T theorems as if they were interpretive—is also hard to spell out further here, given our present understanding of cognition and the brain. One potential way of conceptualizing the assumption is in terms of the familiar idea of the mind as computer. We might imagine the brain as manipulating an internal T theory that takes sentences of a natural language as input and computes T sentences for them in some "brain language" or "language of thought." Suppose that the outputs of these computations are passed to the various mental processors or modules responsible for inference, belief formation, planning of action, and suppose that these modules process the information to infer, form beliefs, plan, etc. The processors that need information about the literal meanings of sentences receive T theorems as their inputs and then proceed under the assumption that the RHSs of these T theorems give the meanings of what's mentioned on the left. In this way, information derived from a theory of truth is treated as information about meaning by the very way the mind passes information around—by the mind's functional architecture, so to speak.[15]

Our third assumption, that speakers acquire an interpretive T theory in the course of development, requires a bit more comment and involves a number of interesting subtleties. First of all, it's clear that this third assumption interacts strongly with the other two. If we are designed to acquire a T theory and to treat any T theory we acquire as interpretive, then we had better acquire the right one. If we somehow came to possess a noninterpretive T theory of French, then by assumption, we would necessarily apply it to the words of French speakers, misunderstanding them, forming wrong views about their beliefs, acting inappropriately toward them, and so on. How is it, then, that we acquire an interpretive T theory, and what *ensures* that the T theory we acquire *is* interpretive?

We suggest that nothing guarantees that the T theories we acquire are interpretive for those around us. Rather, this is a contingent result. That people do quite generally acquire an interpretive T theory is, we suggest, the product of two factors: universal grammar and the very context in which natural language is acquired. In the first chapter we observed that there must be

principles of universal grammar heavily constraining the hypotheses that natural-language learners make about the meanings of words and phrases. A child acquiring *rabbit*, for example, has available as hypotheses not only (26a) but also (26b), (26c), and an indefinite number of others:

(26) a. *Rabbit* refers to rabbits.
b. *Rabbit* refers to undetached rabbit parts.
c. *Rabbit* refers to rabbits, and either dogs are shaggy or dogs are not shaggy.

Similarly, a child acquiring the meaning of the structure $[_{NP}$ NP's N'] has available (27a) but also (27b), (27c), and an indefinite number of other hypotheses:

(27) a. *Rosa's book* refers to the book that Rosa has.
b. *Rosa's book* refers to the book that Rosa is.
c. *Rosa's book* refers to the book that Rosa has, and 2 is an even prime number.

Incorporating (26b, c), (27b, c), or any other equally wrong axioms into the T theory will produce uninterpretive T theorems for sentences containing *rabbit* or the possessive structure *Rosa's book*. Since we assume that UG constraints on lexical and phrasal acquisition play a role in excluding these candidates, UG will serve to guide learners to an interpretive T theory.

A second important factor in acquiring interpretive T theories, we believe, is the very situation in which natural language is acquired. Languages are internalized in the context of communication—in the context of activities such as speaking and understanding. As T theory learners, we hypothesize and fix our rules while engaged in the practical tasks of trying to get our thoughts across to others and trying to grasp what others are trying to say to us. The primary data we use to fix semantic rules in the first place is thus speech produced in communicating with others. Semantic rules are, if you like, fundamentally hypotheses conjectured and tested so as to be interpretive of the speech of others, that is, so as to make sense of their meanings and to make our speech meaningful to them.

What is involved in testing such hypotheses? Briefly put, it seems that we try to see whether interpreting people's speech in the way we conjecture makes sense of their overall motions and interactions with the environment. That is, we try to see whether our interpretation of their speech makes their interactions appear *rational*. In the learning situations described above, for example,

a child might plausibly decide between (26a) and (26c) on these grounds. If sentences containing *rabbit* are interpreted using (26a), then such sentences are simply about rabbits. If they are interpreted using (26c), however, then sentences containing *rabbit* are about rabbits and shaggy dogs. Presumably, a child or other learner would be able to determine by context whether shaggy dogs are under discussion or not, for example, by noting whether the person speaking held up a picture of a dog, made sounds like a dog, or in some other way introduced dogs into the subject matter. By examining the behavior of the speaker and making basic assumptions about the rationality of their actions, we can hypothesize about the meanings of their words. The principles of reasoning we are employing here are, of course, ones that apply not only to speech behavior but also to human action as a whole. We evaluate not only our hypotheses about the meaning of others' speech in this way but also our conjectures about their beliefs, motives, fears, etc. All are judged under the goal of trying to make maximal sense of their actions and behavior.

It is worth stressing, in conclusion, that our answer to the information question—our claims about how knowledge of T theorems might be capable of supporting judgments of meaning—is predicated on the assumption that there is a positive answer to the extension question. It remains a separate hypothesis that an interpretive T theory can actually be given for a full natural language. Our point here is that, on the assumption that such a T theory can be given, knowledge of this theory could account for judgments of actual meaning. In principle, then, judgments of actual meaning are within the scope of what T theories can explain.

2.3 The Theoretical Domain of Semantics

We began chapter 1 with an account of various phenomena constituting the pretheoretical domain of semantics. And we also noted certain important properties that semantic rules should have, such as compositionality. Let us return to these phenomena and see how the T hypothesis accommodates them.

2.3.1 Recasting the Pretheoretical Domain

In discussing the pretheoretical domain of semantics, we said that we have to be ready to adjust our view of the data in the course of articulating a formal theory. We have to be ready to see the data significantly redescribed, or even