UNDERSTANDING UNDERSTANDING:
Semantics, Computation, and Cognition

William J. Rapaport

Department of Computer Science, Department of Philosophy,
and Center for Cognitive Science
State University of New York at Buffalo, Buffalo, NY 14260

rapaport@cs.buffalo.edu
http://www.cs.buffalo.edu/pub/WWW/faculty/rapaport/

DRAFT

March 8, 2007

Contents

List of Figures

The world we perceive is actually an illusion created within our mind, and there instead exists another world beyond our human perceptions, which is accessible only through the powers of the imagination.

—W. R. Hohenberger[1]

You can only look at things from where you stand.

—Kate Bush[2]

The Chinese room shows what we knew all along: syntax by itself is not sufficient for semantics. (Does anyone actually deny this point, I mean straight out? Is anyone actually willing to say, straight out, that they think that syntax, in the sense of formal symbols, is really the same as semantic content, in the sense of meanings, thought contents, understanding, etc.?)

—John Searle[3]

My thesis is that (suitable) purely syntactic symbol-manipulation of a computational natural-language-understanding system's knowledge base suffices for it to understand natural language.

—William J. Rapaport[4]

Does that make any sense? Yes: Everything makes sense. The question is: What sense does it make?

—Stuart C. Shapiro[5]

[1] Quotation on a business card, of unknown origin.

[2] From a radio interview on *All Things Considered*, heard on WBFO–FM, 21 January 1994. This was part of her answer to the question of how a woman could know if she would write differently were she a man. Bush pointed out that she was *not* a man and so *couldn't* know.

[3] Searle 1993: 68.

[4] Rapaport 1988b: 85–86.

[5] In conversation, 19 April 1994.

Chapter 1

SEMANTICS AS CORRESPONDENCE.

1.1 THE FUNDAMENTAL PRINCIPLE OF UNDERSTANDING.

I have heard it said (in connection, as I recall, with quantum mechanics, ascribed to John von Neumann in conversation with Einstein) that you never really understand a new theory—you just get used to it. Taking this as our text, I want to explore its meaning. What is it to understand what someone says, to understand what is expressed in language—to understand language? I suggest the following answer:

The Fundamental Principle of Understanding:

> To understand something is either
>
> 1. to understand it *in terms of something else*, or else
> 2. to "get used to it".

I cannot think of any alternatives that cannot be seen, upon some analysis, to fall under one of these two, admittedly vague (for now!), categories.

Type-1 understanding is *relative*: One understands something relative to one's understanding of another thing. It is a *correspondence theory* of understanding (or of meaning, or of semantics—terms that, for now, I will take as rough synonyms). The correspondence theory of truth is a special case: A sentence about the world is true if and only if it corresponds to (or "matches") the world, where "correspondence" can be explicated à la Tarski.

Type-2 understanding is *non-relative*. 'Absolute' or 'foundational', although plausible alternatives to contrast with 'relative', are too strong. They connote or suggest some sort of "grounding" or "ultimate truth", which is not what I have in mind, though we shall return to the grounding issue. Or, perhaps, type-2 understanding *is* relative—but to itself: To understand something by getting used to it is to understand it in terms of *itself*, perhaps to understand *parts* of it in terms of the *rest* of it. The coherence theory of truth is a special case: A sentence is true if and only if it coheres with the rest of what one takes to be true, where "coherence" can be taken as a kind of relative consistency.

So, both types of understanding can be thought of as relative: type-1 understanding as *externally* relative, type-2 understanding as *internally* relative. Type-1 understanding concerns correspondences between two domains; type-2 understanding concerns syntax.

Since type-1 understanding is relative to the understanding of something *else*, one can only understand something in this first sense if one has *antecedent* understanding of the other thing. How, then, does one understand the other thing? Recursively speaking, either by understanding it relative to some third thing, or by understanding it *in itself*—by being used to it. Either this "bottoms out" in some domain that is understood non-relativistically, or there is a large circle of domains each of which is understood relative to the next. In either case, our understanding bottoms out in "syntactic" understanding of that bottom-level domain or of that large domain consisting of the circle of mutually or sequentially understood domains.

'Correspondence' and 'syntactic understanding' are convenient shorthand expressions that need to be expanded upon. We will examine correspondence first. Before embarking on that, it will be worthwhile to specify a bit more precisely how I will be using the terms 'syntax' and 'semantics'. I will, in fact, be using them in the classic sense due to Charles Morris (1938):

> One may study the relations of signs to the objects to which the signs are applicable. ... [T]he study of this ... will be called *semantics*. Or the subject of study may be the relation of signs to interpreters. ... [T]he study of this dimension will be named *pragmatics*.[1]
>
> One important relation of signs has not yet been introduced: the formal relation of signs to one another. ... [T]he study of this dimension will be named *syntactics*. (Morris 1938: 6–7.)[2]

Thus,

- *Syntax* concerns the relations that symbols have among themselves and the ways in which they can be manipulated.

- *Semantics* concerns the relations between symbols, on the one hand, and the things the symbols "mean", on the other.

Semantics, thus understood, always concerns two distinct domains: a domain of things taken as symbols and governed by rules of syntax, and a domain of other things. These two domains can be called, respectively, 'domain' and 'range', or 'domain' and 'co-domain', or 'syntactic domain' and 'semantic domain'. There must also be a relation between these two domains—the "semantic relation". (For an example of what can go wrong if the relation is not as expected, see Figure 1.1.)

Understanding, in the usual and familiar sense of type-1 understanding, is considered to be a semantic enterprise in this sense of semantics. But even this needs to be examined, because it has some surprising ramifications. Once these are seen, we can turn to the less familiar, type-2 sense of understanding as a syntactic enterprise.

When faced with some new phenomenon or experience, we seek to understand it. Perhaps this need to understand has some evolutionary survival value; perhaps it is uniquely human. Our first strategy in such a case is to find something, no matter how incomplete or inadequate, with which to *compare* the new phenomenon or experience. By thus *interpreting* the "unknown" or "new" in terms of the "known" or "given", we can seek analogies that will begin to satisfy, at least for the moment, our craving for understanding. For instance, I found a recent film, *My Twentieth Century*, to be very confusing (albeit quite entertaining—part of the fun was trying to figure out what it was all about, trying to understand it). I found that I could understand it—at least as a working hypothesis—by mapping the carefree character Lili to the pleasure-seeking, hedonistic aspects of 20th-century life; another character—her serious, twin sister, Dora—to the revolutionary political activist, social-caring aspects of 20th-century life; and the third main character—a professor—to the rational, scientific aspects of 20th-century life. The film, however, is quite complex, and these mappings—these correspondences or analogies—provided for me at best a weak, inadequate understanding. The point, however, is that I *had to*—I was *driven to*—find *something* in terms of which I could make sense of what I was experiencing.

[1] For reasons that will become clearer as we go on, it is arguable that I should be more concerned with pragmatics than with semantics.

[2] For an interesting discussion of the relationships among syntax, semantics, and pragmatics, see Posner 1992.

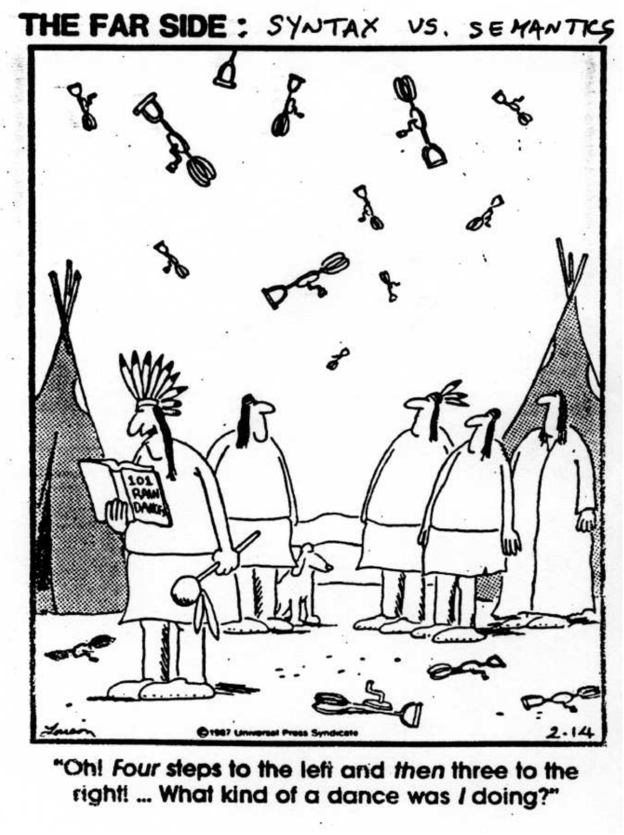

THE FAR SIDE : SYNTAX VS. SEMANTICS
101 RAIN DANCES
©1987 Universal Press Syndicate
2-14
"Oh! *Four* steps to the left and *then* three to the right! ... What kind of a dance was *I* doing?"

Figure 1.1: The relation of syntax to semantics.

Something like this same need for connections as a basis for understanding, as a way to anchor oneself in unchartered waters, can be seen in the epiphenal well-house episode in the life of Helen Keller. With water from the well running over one hand while her teacher Annie Sullivan finger-spells 'w-a-t-e-r' in the other, Helen suddenly understands that 'w-a-t-e-r' means water (Keller 1905). This image of one hand literally in the semantic domain and the other literally in the syntactic domain is striking. By "co-activating" her knowledge (her understanding) of the semantic domain (viz., her experiences of the world around her) and her knowledge of the syntactic domain (viz., her experiences of finger-spellings), she was able to "integrate" (or "bind") these two experiences and thus understand (cf. Mayes 1991: 111). (Or was it as simple as that? We'll return to this celebrated episode in Chapter **??**.)

Before turning to the most well-known and influential theory of semantics as correspondence—Tarski's—let's pause to consider whether there is a sense of semantics other than that of correspondence. After all, many philosophers and linguists look with scorn upon the various mathematical enterprises of formal or model-theoretic semantics. Is there an alternative to this entire enterprise of semantics as a correspondence between two domains? As far as I can tell, there is not. At least, there is not as long as one is willing to talk about "pairings" of sentences (or their structural descriptions) with meaning (cf. Higginbotham 1985: 3). That is, if we are to talk *at all* about "the meaning *of* a sentence", we must be talking about *two* things: sentences and meanings. Thus, there must be two domains: the domain of sentences, described syntactically, and the domain of the semantic interpretation.

There is, however, another kind of semantics, one that linguists not of the formal persuasion study. In this kind of semantics, one is concerned not with what the meanings of linguistic items are, but with semantic relationships among linguistic items: synonymy, implication, etc.[3] These relationships are usually distinct from, though sometimes dependent upon, syntactic relationships. But note that they are, nonetheless, relationships *among linguistic, that is, syntactic, items*. Hence, on our terms, they, too, are "syntactic", not "semantic" (cf. §**??**, below; Kean Kaufmann tells me that *cognitive* linguistics is not to be included here, presumably because it pairs sentences with meanings "in the head" ("cognitive" meanings), in which case, of course, it is a correspondence theory of semantics.) So, semantics is either correspondence or else syntactic.

1.2 TARSKIAN SEMANTICS.

1.2.1 Syntactic Systems.

On the standard view, the syntactic domain is usually some (formal or formalized) language $\mathcal{L}$, which is described syntactically—that is, in terms of its symbols and rules for manipulating them. Thus, for instance, $\mathcal{L}$ might be described as having *terms*, perhaps of two (simple, or atomic) kinds: *individual constants* $a, b, \ldots$ (for example, proper names or other nouns) and *individual variables* $u, v, \ldots$ (for example, pronouns). "New" (complex, or molecular) terms (for example, noun phrases) can be constructed from "old" (whether atomic or molecular) ones by means of *function symbols* of various arities, $f, g, \ldots, f_i, \ldots$ (for example, 'the father of ...', 'the average of ... and __'), together with "grammar" rules specifying the "legal" structure (or "spellings") of such molecular terms (say, if $t_1, \ldots, t_n$ are terms, and f^n is an n-place function symbol, then$\ulcorner f^n(t_1, \ldots, t_n)\urcorner$ is a term). In addition, $\mathcal{L}$ will have *predicate symbols* of various arities, $A, \ldots, Z, A_i, \ldots$ (for example, verb phrases), *connectives* and *quantifiers*, $\neg, \vee, \forall, \ldots$ (for example, 'it is not the case that ...', '... or __', 'for all ..., it is the case that __'), and more "grammar" rules specifying the "legal" structure of *well-formed formulas* (or sentences): If $t_1, \ldots, t_n$ are terms, and P^n is an n-place predicate symbol, then $\ulcorner P^n(t_1, \ldots, t_n)\urcorner$ is a well-formed formula (wff); if φ and ψ are wffs, and v is an individual variable, then$\ulcorner \neg\psi\urcorner$,$\ulcorner(\varphi \vee \psi)\urcorner$,$\ulcorner \forall v[\varphi]\urcorner$ are wffs.

Note that $\mathcal{L}$ is a *language*. Sometimes $\mathcal{L}$ is augmented with a *logic*: Certain wffs of $\mathcal{L}$ are distinguished as *axioms* (or "primitive *theorems*"), and *rules of inference* are provided that specify how to produce "new" theorems from "old" ones. For instance, if φ and$\ulcorner(\varphi \rightarrow \psi)\urcorner$ are theorems, then so is ψ. A *proof* of a wff ψ

[3] I am indebted to Kean Kaufmann and Matthew Dryer for helping me to see this.

(from a set of wffs Σ) is a sequence of wffs ending with ψ such that every wff in the sequence is either an axiom (or a member of Σ) or follows from previous wffs in the sequence by one of the rules of inference.

And so on. I will assume that the reader is familiar with the general pattern (see, for example, Rapaport 1992ab for more details). The point is that all we have so far are symbols and rules for manipulating them either linguistically (to form wffs) or logically (to form theorems). All we have so far is syntax in Morris's sense. Actually, in my desire to make the example perspicuous, I may have given you a misleading impression by talking of "language" and "logic", of "nouns" and "verb phrases", etc. For such talk tends to make people think either that I *was* talking, albeit in a very strange way, about language and nouns and verbs—good old familiar languages like English with nouns and verbs like 'dog' and 'run'—or that I had that in the back of my mind as an intended interpretation of the symbols and rules. But what I intend by 'symbols' are just marks, (perhaps) physical inscriptions or sounds, that have only some very minimal features such as having distinguished, relatively unchanging shapes capable of being recognized when encountered again.

So, let me offer a somewhat less familiar syntactic domain $\mathcal{L}'$, which I will call this time, not a "language", but merely a "symbol system". First, I need to show you the symbols of $\mathcal{L}'$. To really make my point, these should be quite arbitrary, say, boxes, circles, squiggles of various kinds. But I will make life a bit easier for the reader and the typesetter by using letters and numerals.

$\mathcal{L}'$ consists of the following symbols:

$A_1, \ldots, A_i, \ldots;$
$F_0, F_1, F_2, F_3;$
(,), ,, ;; [i.e., a left-parenthesis, a right-parenthesis, a comma, and a semi-colon]
R

I want to show you a certain class K of symbols of $\mathcal{L}'$. To talk about them, I'll need another set of symbols that are not part of $\mathcal{L}'$, so we'll let 'A', 'B', 'C', 'B_1', 'B_2', ... be variables ranging over the members of K. Now, here are the members of K:

1. $A_1, \ldots, A_i, \ldots \in K$
2. If $A, B \in K$, then$\ulcorner F_0(A)\urcorner, \ulcorner F_1(A,B)\urcorner, \ulcorner F_2(A,B)\urcorner, \ulcorner F_3(A,B)\urcorner \in K$.
3. Nothing else is in K.

We could ask questions of this formal symbol system. For instance, which molecular symbols are in K? By suitable symbol manipulation, following (1)–(3), we can ascertain that $A_1, A_{100}, F_0(A_{100}), F_0(F_0(A_{100})), F_3(F_0(F_0(A_{100})), F_2(A_1, A_{100})) \in K$, but that $F_0(F_0), B \notin K$.

Now, let's make $\mathcal{L}'$ a bit more interesting. Let $H \subseteq K$; let $A, B \in K$; and let's say that an (H, A)-*sequence* is a sequence of members of K such that A is the last item in the sequence, and, if B is in the sequence, then either $B \in H$ or there is a set $\{B_1, \ldots, B_n \mid (\forall 1 \leq i \leq n)[B_i \in K]\}$ such that$\ulcorner R(B_1, \ldots, B_n; B)\urcorner \in \mathcal{R}$, where $\mathcal{R}$ is defined as follows (remember that 'R' is a symbol of $\mathcal{L}'$; I am defining $\mathcal{R}$ as consisting of certain sequences of symbols beginning with 'R'):

$\mathcal{R}$1. $\ulcorner R(A; F_1(A,B))\urcorner \in \mathcal{R}$

$\mathcal{R}$2. $\ulcorner R(B; F_1(A,B))\urcorner \in \mathcal{R}$

$\mathcal{R}$3. $\ulcorner R(F_1(A,B), F_0(A); B)\urcorner \in \mathcal{R}$

$\mathcal{R}$4. $\ulcorner R(F_1(A,B), F_0(B); A)\urcorner \in \mathcal{R}$

$\mathcal{R}$5. $\ulcorner R(F_2(A,B); A)\urcorner \in \mathcal{R}$

$\mathcal{R}$6. $\ulcorner R(F_2(A,B); B)\urcorner \in \mathcal{R}$

$\mathcal{R}$7. $\ulcorner R(A, B; F_2(A, B))\urcorner \in \mathcal{R}$

$\mathcal{R}$8. $\ulcorner R(F_3(A, B), A; B\urcorner \in \mathcal{R}$

$\mathcal{R}$9. If there is an (H, B)-sequence whose first item is A,
then$\ulcorner R(; F_3(A, B))\urcorner \in \mathcal{R}$ [Note: There is no symbol between '(' and ';'.]

$\mathcal{R}$10. If there is an $(H, \ulcorner F_2(B, F_0(B))\urcorner)$-sequence whose first item is A,
then$\ulcorner R(; F_0(A))\urcorner \in \mathcal{R}$

$\mathcal{R}$11. If there is an $(H, \ulcorner F_2(B, F_0(B))\urcorner)$-sequence whose first item is $F_0(A)$,
then$\ulcorner R(; A)\urcorner \in \mathcal{R}$

$\mathcal{R}$12. Nothing else is in $\mathcal{R}$.

We can now ask more questions of our system. For instance, which symbols A are such that$\ulcorner R(; A)\urcorner \in \mathcal{R}$? By suitable symbol manipulations, following $\mathcal{R}$1–$\mathcal{R}$12, we can ascertain that, for example, $R(; F_3(A_0, A_0)) \in \mathcal{R}$ (this is actually fairly trivial, since $\langle A_0 \rangle$ is an (A_0, A_0)-sequence whose first item is A_0).

Hard to read, isn't it! You feel the strong desire to try to understand these squiggles, don't you? You would probably feel better if I showed you some other domain with which you were more comfortable, more familiar, into which you could map these squiggles. I will. But not yet. Of course, I could be sadistic and suggest that you "get used to" $\mathcal{L}'$ by manipulating its symbols and learning more about the members of K and $\mathcal{R}$. You could do that, and you *would* learn more. But I won't be that mean. First, we need to move away from pure syntax and find out what semantics consists of.

1.2.2 Semantic Interpretations.

Given some syntactic domain—some formal symbol system—one can ask two sorts of questions about it. The first sort is exemplified by those we asked above: What are the members of K? Of $\mathcal{R}$? These are purely "internal", syntactic, questions. The second sort is, in short: What's the meaning of all this? What do the symbols mean (if anything)? What, for example, is so special about the members of K or the symbols of the form$\ulcorner R(; A)\urcorner$? To answer this sort of question, we must go outside the syntactic domain: We must provide "external" entities that the symbols mean, and we must show the mappings—the associations, the correspondences—between the two domains.

Now, a curious thing happens: I need to show you the semantic domain. If I'm very lucky, I can just point it out to you—we can look at it together, and I can describe the correspondences ("The symbol A_{37} means that red thing over there."). But, more often, I have to describe the semantic domain to you in ... symbols, and hope that the meaning of *those* symbols will be obvious to you. (We'll return to this problem in §1.7).

As an example, let's see how to provide a semantic interpretation of our first formal symbol system, $\mathcal{L}$. Since $\mathcal{L}$ had individual terms, function symbols, and predicate symbols—which could be combined in various (but not arbitrary) ways—I need to provide meanings for each such symbol as well as for their legal combinations. So, we'll need a non-empty set **D** of things that the terms will mean—a **D**omain of interpretation (sometimes called a **D**omain, or universe, of discourse)—and sets **F** and **R** of things that the function and relation symbols will mean, respectively. These three sets can be collectively called **M** (for **M**odel). What's in **D**? Well, anything you want to talk or think about. What are in **F** and **R**? Functions and relations on **D** of various arities—that is, anything you want to be able to say about the things in **D**. That's our *ontology*, what there is.

Now for the correspondences. To say what a symbol of $\mathcal{L}$ means in **M** (what the meaning, from **M**, of a symbol of $\mathcal{L}$ is), we can define an interpretation function $I : \mathcal{L} \rightarrow \mathbf{M}$ that will assign to each symbol of $\mathcal{L}$ something in **M** (or it might be an interpretation *relation* if we wish to allow for ambiguity), as follows:

1. If t is an individual term of $\mathcal{L}$, then $I(t) \in \mathbf{D}$.
   (*Which* element of $\mathbf{D}$? Whichever you want, or, if we spell out $\mathcal{L}$ and $\mathbf{D}$ in more detail, I'll tell you; for example, perhaps I('William J. Clinton') = the 42nd President of the U.S., if 'William J. Clinton' is an individual constant of $\mathcal{L}$, and $\mathbf{D}$ is the set of humans.)

2. If f is a function symbol of $\mathcal{L}$, then $I(f) \in \mathbf{F}$.

3. If $\ulcorner f(t_1, \ldots, t_n)\urcorner$ is a (molecular) term of $\mathcal{L}$,
   then $I(\ulcorner f(t_1, \ldots, t_n)\urcorner) = I(f)(I(t_1), \ldots, I(t_n)) \in \mathbf{D}$.
   (I.e., the interpretation of $\ulcorner f(t_1, \ldots, t_n)\urcorner$ will be the result of applying (a) the function that is the interpretation of f to (b) the elements of $\mathbf{D}$ that are the interpretations of the t_i; and the result will be an element of $\mathbf{D}$.)

4. If P is a predicate symbol of $\mathcal{L}$, then $I(P) \in \mathbf{R}$.

So far, so good. Now, what do wffs mean? Those philosophers and logicians who take n-place functions and relations to be ordered n-tuples—functions and relations "in extension"—tend to talk about "truth values" of wffs rather than "meanings". Others, who take functions and relations "in intension" can talk about the meanings of wffs as being "states of affairs" or "situations" or "propositions", variously defined. I, myself, fall in the latter camp, but for the sake of simplicity of exposition, I'll go the other route for now. Continuing, then, we have:

5. If φ is a wff, then $I(\varphi) \in \{0, 1\}$, where, intuitively, we'll say that φ is "true" if $I(\varphi) = 1$ and that φ is "false" if $I(\varphi) = 0$. In particular, where P is an n-place predicate symbol, $t_1, \ldots, t_n$ are terms, v is an individual variable, and φ, ψ are wffs:
   (a) $I(\ulcorner P(t_1, \ldots, t_n)\urcorner) = 1$ iff $\langle I(t_1), \ldots, I(t_n)\rangle \in I(P)$.
   (b) $I(\ulcorner \neg\varphi\urcorner) = 1$ iff $I(\varphi) = 0$
   (c) $I(\ulcorner(\varphi \vee \psi)\urcorner) = 1$ iff $I(\varphi) = 1$ or $I(\psi) = 1$ (or both)
   (d) $I(\ \ulcorner\forall v[\varphi]\urcorner) = 1$ iff $I'(\varphi) = 1$ for every I' that differs from I at most on what I' assigns to v.

Now, what kind of function is I? Clearly, it is a homomorphism; that is, it satisfies a principle of compositionality: The interpretation of a molecular symbol is determined by the interpretations of its atomic constituents in the manner spelled out above. In the ideal case, I is an *isomorphism*—a 1–1 and onto homomorphism; that is, *every* item in $\mathbf{M}$ is the meaning of *just one* symbol of $\mathcal{L}$. (Being onto is tantamount to $\mathcal{L}$'s being "complete". Perhaps isomorphism is less than ideal, at least for the case of natural languages. David P. Wilkins (1995: 381) has observed that when one studies, not isolated or made-up sentences, but

> ... real, contextualised utterances ... it is often the case that all the elements that one would want to propose as belonging to semantic structure have no overt manifestations in syntactic structure. ... [T]he degree of isomorphism between semantic and syntactic structure is mediated by pragmatic and functional concerns ....

In this ideal situation, $\mathbf{M}$ is a virtual duplicate or mirror image of $\mathcal{L}$. (Indeed, $\mathbf{M}$ could *be* $\mathcal{L}$ itself (cf. Chang & Keisler 1973: 4ff), but that's not very interesting or useful for *understanding* $\mathcal{L}$.) In less ideal circumstances, there might be symbols of $\mathcal{L}$ that are *not* interpretable in $\mathbf{M}$; in that case, I would be a *partial* function. Such is the case when $\mathcal{L}$ is English and $\mathbf{M}$ is the world ('unicorn' is English, but unicorns don't exist), though if we "enlarge" or "extend" $\mathbf{M}$ in some way, for example, if we take $\mathbf{M}$ to be Meinong's *Aussersein* instead of the actual world, then we can make I total (cf. Rapaport 1981). In another less ideal circumstance, "Horatio's Law" might hold: There are more things in $\mathbf{M}$ than in $\mathcal{L}$; that is, there are elements of $\mathbf{M}$ not expressible in $\mathcal{L}$: I is not onto. And, as noted earlier, I might be a relation, not a function, so $\mathcal{L}$

would be ambiguous. There is another, more global, sense in which $\mathcal{L}$ could be ambiguous: By choosing a different $\mathbf{M}$ (and a different I), we could give the symbols of $\mathcal{L}$ entirely distinct meanings. Worse, the two $\mathbf{M}$s need not be isomorphic. (This can happen in at least two ways. First, the cardinalities of the two $\mathbf{D}$s could differ. Second, suppose $\mathcal{L}$ is a language for expressing mathematical group theory. Then $\mathbf{M}_1$ could be an infinite cyclic group (for example, the integers under addition), and $\mathbf{M}_2$ could be $\mathbf{M}_1 \times \mathbf{M}_1$, which, unlike $\mathbf{M}_1$, has two disjoint subgroups (except for the identity).[4]

Let's consider an example in detail; I'll tell you what the symbols of $\mathcal{L}'$ mean. First, I need to show you $\mathbf{M}$. To do that, I need to show you $\mathbf{D}$: $\mathbf{D}$ will include the *symbols*: $\varphi_1, \ldots, \varphi_i, \ldots$ (so, I'm explaining one set of symbols in terms of another set of symbols; be patient). $\mathbf{D}$ will also include these symbols: $\neg, \vee, \wedge, \rightarrow$. Now I can tell you about K (in what follows, let A_i be the ith atomic symbol of K, let φ_i be the ith atomic symbol of $\mathbf{D}$, and let $A, B \in K$):

$$\begin{array}{l}
I(A_i) = \varphi_i \\
I(F_0) = \neg \\
I(F_1) = \vee \\
I(F_2) = \wedge \\
I(F_3) = \rightarrow \\
I(\ulcorner F_0(A)\urcorner) = \ulcorner \neg I(A)\urcorner \\
I(\ulcorner F_1(A,B)\urcorner) = \ulcorner (I(A) \vee I(B))\urcorner \\
I(\ulcorner F_2(A,B)\urcorner) = \ulcorner (I(A) \wedge I(B))\urcorner \\
I(\ulcorner F_3(A,B)\urcorner) = \ulcorner (I(A) \rightarrow I(B))\urcorner
\end{array}$$

I assume, of course, that you know what '$\neg$',$\ulcorner(I(A) \rightarrow I(B))\urcorner$, etc., are (namely, the negation sign, a material conditional wff, etc.). So, the elements of K are just wffs of propositional logic (as if you didn't know)! What about $\mathcal{R}$? Well: $I(R) = \vdash \in \mathbf{R}$ (where $\mathbf{R}$, of course, is part of $\mathbf{M}$); that is, R means the deducibility relation on wffs of propositional logic. So, the elements of $\mathcal{R}$ are rules of inference:

$$\begin{array}{lcl}
I(\ulcorner R(A; F_1(A,B))\urcorner) & = & A \vdash \ulcorner(A \vee B)\urcorner \text{ (that is, }\vee\text{-introduction)} \\
I(\ulcorner R(B; F_1(A,B))\urcorner) & = & B \vdash \ulcorner(A \vee B)\urcorner \text{ (that is, }\vee\text{-introduction)} \\
I(\ulcorner R(F_1(A,B), F_0(A); B)\urcorner) & = & \ulcorner(A \vee B)\urcorner, \ulcorner \neg A\urcorner \vdash B \text{ (that is, }\vee\text{-elimination)} \\
I(\ulcorner R(F_1(A,B), F_0(B); A)\urcorner) & = & \ulcorner(A \vee B)\urcorner, \ulcorner \neg B\urcorner \vdash A \text{ (that is, }\vee\text{-elimination)} \\
I(\ulcorner R(F_2(A,B); A)\urcorner) & = & \ulcorner(A \wedge B)\urcorner \vdash A \text{ (that is, }\wedge\text{-elimination)} \\
I(\ulcorner R(F_2(A,B); B)\urcorner) & = & \ulcorner(A \wedge B)\urcorner \vdash B \text{ (that is, }\wedge\text{-elimination)} \\
I(\ulcorner R(A, B; F_2(A,B))\urcorner) & = & A, B \vdash \ulcorner(A \wedge B)\urcorner \text{ (that is, }\wedge\text{-introduction)} \\
I(\ulcorner R(F_3(A,B), A; B)\urcorner) & = & \ulcorner(A \rightarrow B)\urcorner, A \vdash B \text{ (that is, }\rightarrow\text{-elimination, or Modus Ponens)}
\end{array}$$

Before we can finish interpreting R, I need to tell you what an (H, A)-sequence means: It is a proof of $I(A)$ from hypotheses $I(H)$ (where, to be absolutely precise, I should specify that, where $H = \{A, B, \ldots\} \subseteq K, I(H) = \{I(A), I(B), \ldots\}$). So:

$I(\mathcal{R}9)$ is:
if there is a proof of $I(B) \in \mathbf{D}$ from a set of hypotheses $I(H)$ whose first line is $I(A)$, then
$\vdash \ulcorner(I(A) \rightarrow I(B))\urcorner$
(that is, $\rightarrow$-introduction, or Conditional Proof)

$I(\mathcal{R}10)$ is:
if there is a proof of $\ulcorner(I(B) \wedge \neg I(B))\urcorner$ from a set of hypotheses $I(H)$ whose first line is $I(A)$, then
$\vdash \ulcorner \neg I(A)\urcorner$
(that is, $\neg$-introduction)

[4] I am grateful to Nicolas Goodman for this example.

$I(\mathcal{R}11)$ is:
if there is a proof of $\ulcorner(I(B) \wedge \neg I(B))\urcorner$ from a set of hypotheses $I(H)$ whose first line is $\ulcorner\neg I(A)\urcorner$, then
$\vdash I(A)$
(that is, $\neg$-elimination)

So, now you know: $\mathcal{L}'$ is just ordinary propositional logic in a weird notation. Of course, I could have told you what the symbols of $\mathcal{L}'$ mean in terms of a *different* model $\mathbf{M}'$, where $\mathbf{D}'$ consists of states of affairs and Boolean operations on them. In that case, $\mathcal{L}'$ just *is* ordinary propositional logic. That is, $\mathbf{M}$ is itself a syntactic formal symbol system (namely, $\mathcal{L}$!) whose meaning can be given in terms of $\mathbf{M}'$, but $\mathcal{L}'$'s meaning can be given either in terms of $\mathbf{M}$ *or* in terms of $\mathbf{M}'$.

There are several lessons to be learned from this. First, $\mathcal{L}'$ is not a very "natural" symbol system. Usually, when one presents the syntax of a formal symbol system, one already has a semantic interpretation in mind, and one *designs* the syntax to "capture" that semantics: In a sense that will become clearer in the next section, the syntax is a model—an implementation—of the semantics.

Second, it is possible and occasionally even useful to allow *one syntactic* formal symbol system to be the semantic interpretation of *another*. Of course, this is only useful if the interpreting syntactic system is antecedently understood. How? In terms of *another* domain with which we are antecedently familiar! So, in our example, the unfamiliar $\mathcal{L}'$ was interpreted in terms of the more familiar $\mathbf{M}$ (i.e., $\mathcal{L}$), which, in turn, was interpreted in terms of $\mathbf{M}'$. And how is it that we understand what states of affairs in the world are? Well . . . we've just gotten used to them.

Finally, note that $\mathbf{M}$ in our example is a sort of "swing" domain: It serves as the *semantic* domain relative to $\mathcal{L}'$ and as the *syntactic* domain relative to $\mathbf{M}'$. We can have a "chain" of domains, each of which except the first is a semantic domain for the one before it, and each of which except for the last is a syntactic domain for the one following it. To understand any domain in the chain, we must be able to understand the "next" one. How do we understand the last one? Syntactically. But I'm getting ahead of myself. Let's first look at these "chains" and their possible components.

1.3 THE CORRESPONDENCE CONTINUUM: DATA.

Let's begin with examples—lots of them. The more examples I can show you—the more data there are—then the more you will come to see what I see, to accept my hypothesis (to be stated below). (The examples are summarized in Tables 1.1 and 1.2.) I am going to present you with *pairs* of things: One member of each pair plays the role of the syntactic domain; the other plays the role of the semantic domain. (We'll return to many of them in detail later.)

1. The first example is the obvious one: our old friends $\mathcal{L}'$ and $\mathbf{M}$ (or $\mathbf{M}$ and $\mathbf{M}'$).

2. The next examples come from what I'll call *The Muddle of the Model in the Middle* (cf. Wartofsky 1966). There are two notions of "model" in science and mathematics: We speak of a "mathematical model" of some physical phenomenon, by which we mean a mathematical, usually formal, theory of the phenomenon. In this sense of 'model', a model is a *syntactic* item whose intended semantic interpretation is the physical phenomenon being "modeled". But we also speak of a semantic interpretation of a syntactic domain as a "model", as in the phrase 'model-theoretic semantics'. In this sense of 'model', a model is a *semantic* domain. So we have the following syntax/semantics pairs:

   data/formal theory (that is, theory as interpretation of the data),
   formal theory/set-theoretic (or *mathematical*) *model* (that is, a model of the theory),
   set-theoretic (or *mathematical*) *model/real-world phenomenon*.

	role of the syntactic domain	**role of the semantic domain**
1.	a formal language $\mathcal{L}$	a model $\mathbf{M}$
2.	data	formal theory accounting for the data
	formal theory	set-theoretic model of the theory
	set-theoretic model	real-world phenomenon
3.	caption	newspaper photo
4.	digitized image of handwritten word	word
5.	musical score	performance of the score
6.	play script	performance of the play
7.	novel	movie or play based on the novel
8.	narrative (text)	story told by the narrative
9.	narrative (text)	mental model constructed by reader
10.	(see Table 1.2)	
11.	linguistic or perceptual input	mental model
12.	mental model	actual world
13.	SNePS nodes	concepts (Meinongian objects in Aussersein)
14.	concepts, Meinongian objects	Sein-correlates (Rapaport 1978)
15.	discourse	discourse representation structures
	discourse representation structures	actual world
	discourse	actual world
16.	English text	French translation
	French translation	English text
17.	linguistic expressions	ideas
18.	speech, sign languages	language
19.	map	Earth
20.	blueprint	house
21.	scale model	thing modeled
22.	representational painting	real world
	real world	representational painting
23.	specifications	computer program
24.	computer program	computer process
25.	bits in a computer	data structure
26.	formulas of analysis	geometry
	chaotic systems	continued fractions
	continued fractions	chaotic systems
27.	expressions of language	mentalese tokens
	mentalese tokens	(other) mentalese tokens
	mentalese tokens	designations in world of discourse

Table 1.1: Syntactic and semantic domains.

$$\text{narrative} \rightarrow \text{play} \rightarrow \left\{ \begin{array}{l} \text{opera} \rightarrow \text{ballet}_1 \rightarrow \text{film}_1 \rightarrow \text{novelization}_1 \\ \text{ballet}_2 \\ \text{film}_2 \rightarrow \text{novelization}_2 \rightarrow \text{film}_3 \\ \text{symphony} \rightarrow \text{performances} \end{array} \right.$$

Table 1.2: Example 10. A correspondence continuum. Each syntactic–semantic pair is of the form: syntactic domain → semantic domain, where the latter is an artwork "based on" the former.

The latter, when you think of it, is closely related to—if not identical with—the data that we began with, giving us a cycle of domains! (Cf. Rosenblueth & Wiener 1945: 316.)

3. A *newspaper photograph* can be thought of as a semantic interpretation of its *caption*. There's more, since a cognitive agent who reads the caption and looks at the photo makes further correspondences. For instance, (a) there will be a mental model of the caption—the reader's semantic interpretation of the caption-as-syntax; (b) there will be a mental model of the photo—the reader's semantic interpretation of the photo-as-syntax; and, (c) depending on one's theory of how such picture+caption units are processed, (i) there may be correspondences between these two mental models, or (ii) there may be a single mental model that collates the information from each of these and which, in turn, is a semantic interpretation of the picture+caption unit. (See Srihari & Rapaport 1989, 1990; Srihari 1991ab, 1993ab. In these, option (cii) is taken.)

4. The problem of handwritten and printed word recognition (one of the earliest AI problems—not to mention one of the first tackled by a philosopher (Sayre 1973)) can be approached as follows:

   > Given a digitized image of a word and a lexicon containing the word, produce a ranking of the lexicon such that the word in the image is ranked as close to the top as possible. (Ho 1990.)

   Here, the syntactic domain is the digitized image of a printed or written word (a token), and the semantic domain is the word (a type). The word-recognition system will understand what word it is by providing a semantic interpretation from a lexicon. Note that it does this by pattern matching (or pattern "recognition"): Given a symbol, recognize its pattern (its structure)—that is, classify it.

5. A *musical score*, say, Bach's *Goldberg Variations*, is a piece of syntax; a *performance of* it is a semantic interpretation. And, of course, there could be a performance of the *Goldberg Variations* on piano or on a harpsichord (or even on a synthesizer, a banjo, or a kazoo). For instance, a piano transcription of a symphony is a semantic interpretation of the symphony (cf. Pincus 1990; conversely, Brian Cantwell Smith (1985: 636) considers "musical scores as models of a symphony".)

6. Similarly, the *script* of a play is syntax; a *performance* of the play is a semantic interpretation. For a performance to be a semantic interpretation of the script, an actual *person* would—literally(?)—play the role (that is, be the semantic interpretation) of a *character* in the play. (And Olivier's interpretation of Hamlet is very different from Burton's.) (Scripts are like computer programs; performances are like computer processes; see example 24 and cf. Rapaport 1988.)

7. A *movie* or *play* based on a *novel* can be considered a semantic interpretation of the text. In this case, there must be correspondences between the characters, events, etc., in the book and the play or movie, with some details of the book omitted (for lack of time, say) and some things in the play or movie added (decisions must be made, say, about the colors of costumes, which might not have been specified in the book, just as one can *write* about a particular elephant without specifying whether it's facing left or right, but one can't *show*, *draw*, or *imagine* the elephant without so specifying).

8. Consider a narrative text as a piece of syntax: a certain sequence of sentences and other expressions in some natural language. The *narrative* tells a *story*—the story is a semantic interpretation of the text. On this way of viewing things, the narrative has a "plot"—descriptions of certain events in the story, but not necessarily ordered in the chronological sequence that the events "actually" occurred in. Thus, one story can be told in many ways, some more interesting or suspenseful than others. The story takes place in a "story world". Characters, places, times, etc., in the story world correspond to linguistic descriptions or expressions of them in the narrative. "The" story world in which the events take place need not be unique, since (as in example 7) the narrative need not (indeed, *cannot*, be fully explicit (thus, for example, in one story world corresponding to *The Hound of the Baskervilles*, Sherlock Holmes has a mole on his left arm; in another, he doesn't). The story world as thus described is somewhat of an abstraction. Alternatively, it could be the author's mental model (model of what?)—a structure in the

author's mind, perhaps expressed in his or her language of thought, which the author then expresses as a narrative in natural language. (Cf. Segal 1995.)

9. There is also the reader of the narrative who constructs a mental model of the narrative as he or she reads it. This mental story is a semantic interpretation of the syntactic narrative. Or one could view it as a *theory* constructed from the narrative-as-data (cf. Bruder et al. 1986; Rapaport, Segal, Shapiro, Zubin, Bruder, Duchan, et al. 1989; Rapaport, Segal, Shapiro, Zubin, Bruder, Duchan, & Mark 1989; Duchan et al. 1995).

10. In fact, examples 5–9 suggest a tree of examples: Some *narrative text* might be interpreted as a *play*, on which an *opera* is based. There could be a *film* of a *ballet* based on the *opera*, and these days one could expect a "*novelization*" of the film. Of course, a (different) ballet could be based directly on the play, or a film could have been based directly on the play, then novelized, then re-filmed. Or a symphony might have been inspired by the play, which symphony, of course, will have several performances. And so on. Vincent Canby (1994) calls such a sequence "the usual evolutionary process by which our popular entertainment grows .... A book is turned into a play, the play into a movie, the movie into a stage musical, the stage musical into a movie musical. That's the end of the line, unless the original property somehow becomes a television series." And Budd Schulberg (1995: H5) notes that *On the Waterfront* was "First a Movie, Then a Novel, Now a Play".

11. The *linguistic and perceptual "input"* to a cognitive agent can be considered as a syntactic domain whose semantic interpretation is provided by the agent's *mental model* of his or her (or its) sensory input. (The mental model is the agent's "theory" of the sensory "data"; cf. examples 2 and 9.)

12. The *mental model*, in turn, can be considered as a syntactic language of thought whose semantic interpretation is provided by the *actual world*. In this sense, a person's beliefs are true to the extent that they correspond to the world.

13. Turning to computational "models", and related to example 12, *SNePS nodes* (more generally, terms and expressions of an intensional language of thought) are syntactic items interpretable in terms of *concepts* in a Meinongian Aussersein (cf. Shapiro & Rapaport 1987, 1991).

14. And these *concepts* (Meinongian objects in general) are in turn the syntactic domain for the semantic domain consisting of "*Sein-correlates*"—that is, actual objects in the real world. (Cf. Rapaport 1978.)

15. In Discourse Representation Theory, there is a discourse (which is a linguistic text—a piece of syntax), a (sequence of) discourse representation structures, and the actual world (or a representation thereof), and mappings from the discourse to the discourse representation structures, from the discourse to the world, and from the discourse representation structures to the world. Each such mapping is a semantic interpretation. (See the references in §??. Cf. examples 8 and 9, where the discourse is to the narrative as the discourse representation structure is to the mental model as the world is to the story world. And, of course, one can consider the correspondences, if any, between the story world and the actual world; these, too, are semantic.) (Cf. examples 11 and 12.)

16. A *French translation* of an *English text* can be seen from the point of view of the French speaker as a semantic interpretation of the English syntax. Equally, it can be seen from the point of view of the English speaker as a syntactic expression of the English (cf. Gracia 1990: 533).

17. In general, of course, expressions of a language (words, sentences, etc.) are syntactic; the ideas they express are semantic. (Cf. Harris 1987.)

18. Similarly, language can be considered a semantic domain that can be expressed syntactically in speech and in sign (and in *many ways* in speech (English, French, etc.) and in sign (ASL, BSL, etc.)):

> The idea [about ASL] that language didn't have to be spoken was completely novel [to me]. It meant that language was a capacity of the brain, and if it didn't come out one way, it would come out another. (Harlan Lane, in Coughlin 1991.)

19. The *Earth* is the semantic domain for a global *map*.

20. A *house* is a semantic interpretation of a *blueprint* (cf. Potts 1973, Rapaport 1978, Smith 1985).

21. A *scale model* (say, of an airplane) corresponds to the *thing modeled* (say, the airplane itself) as syntax to semantic interpretation. And, of course, the thing modeled could itself be a scale model, say, a statue; so I could have a model of a statue, which is, in turn, a model of a person. (Cf. Smith 1985, Shapiro & Rapaport 1991).

22. Representational painting provides a syntactic domain corresponding to the real-world semantic domain. Conversely, representational art can be considered a model of the world—a theory of what the world is or looks like.

23. The *specifications* for a computer program—*what* the program is supposed to do—are interpreted by the *program*—which explicates *how* things are done. (Cf. Smith 1985: 640.)

24. A computer *program*, as noted earlier, is a static piece of syntax; a computer *process* can be thought of as its semantic interpretation. And, according to Smith, one of the concerns of knowledge representation is to interpret *processes* in terms of the actual world: "It follows that, in the traditional terminology, the *semantic* domain of traditional programming language analyses [which "take ... semantics as the job of mapping programs onto processes"] should be the knowledge representer's so-called *syntactic* domain" (Smith 1987: 15; cf. p. 17, and p. 18, figs. 7–8).

25. A *data structure* (such as a stack or a record) provides a semantic interpretation of (or, a way of categorizing) the otherwise inchoate and purely syntactic *bits* in a computer:

    > The concept of information in computer science is similar to the concepts of point, line, and plane in geometry—they are all undefined terms about which statements can be made but which cannot be explained in terms of more elementary concepts. ... The basic unit of information is the *bit* .... (Tenenbaum & Augenstein 1981: 1.)
    >
    > [I]nformation itself has no meaning. Any meaning can be assigned to a particular bit pattern as long as it is done consistently. It is the interpretation of a bit pattern that gives it meaning. ... A method of interpreting a bit pattern is often called a *data type*. (Tenenbaum & Augenstein 1981: 6.)
    >
    > Any type in Pascal may be thought of as a pattern or a template. By this we mean that a type is a method for interpreting a portion of memory. When a variable identifier is declared as being of a certain type, we are saying that the identifier refers to a certain portion of memory and that the contents of that memory are to be interpreted according to the pattern defined by the type. The type specifies both the amount of memory set aside for the variable and the method by which that memory is interpreted" (Tenenbaum & Augenstein 1981: 45.)
    >
    > A *data type* is an interpretation applied to a string of bits (Schneiderman 1993: 411.)

    This can be further elaborated: Suppose we have a computer program intended to model the behavior of customers lining up at a bank. Some of the data structures of this program will represent customers. This gives rise to the following transitive syntax–semantics chain: syntactic bits are semantically interpreted by data structures, which, in turn, are semantically interpreted as customers. (For a related, though slightly different view, consider Smith 1982b: 11: "... the notion *program* is inherently defined as a set of expressions whose (Φ-)semantic domain includes *data structures* .... In other words, in a computational process that deals with finance, say, the *general* data structures will likely designate individuals and money and relationships among them, but the terms in the part of the process called a *program* will not designate these people and their money, but will instead designate *the data structures that designate people and money* ...".)

26. As the mathematician N. Steenrod has observed,

> Two views of the same thing reinforce each other. Most of us are able to remember the multitudinous formulas of analysis mainly because we attach to each a geometric picture that keeps us from going astray. (Steenrod 1967: 777.)

That is, the *geometry* is a semantic interpretation of the syntactic *formulas of analysis*. (There is more to say about this, however: For if the syntactic formulas of analysis and the semantic geometry are "two views of the same thing", what is that thing? Perhaps *it* is a semantic domain for which *both* the geometry *and* the analysis are syntactic expressions.)

Similarly, the first sentence of R. M. Corless's "Continued Fractions and Chaos" (1992) begins thusly:

> This paper is meant for the reader who knows something about continued fractions, and wishes to know more about the theory of chaotic systems;[1] (p. 203),

at which point Corless's footnote 1 informs us that

> One referee has remarked that "This describes the referee, who admits to having found the paper interesting. Though, I suspect, now, more people know about chaos than continued fractions." The author is inclined to agree, and hopes that this paper will interest some of these people in continued fractions.

27. Brian Cantwell Smith (1982b: 10–11) considers (a) a mapping Θ from a syntactic *language* or "notational" system to "internal elements"—for example, from words to mentalese tokens—(b) a mapping Φ from the internal elements to "designations" in "the world of discourse", and (c) a mapping Ψ from internal elements to other internal elements, all of which is taken to be "semantical", even the clearly "syntactic" Ψ. And he speaks of "a general significance function ... that recursively specifies Ψ and Φ together ...." (Θ and Φ are reminiscent of William A. Woods's "linguistic semantics" and "philosophical semantics" (Woods 1975: 38–39). David Lewis (1972) has argued that Θ-like mappings are *not* semantic.)

No doubt you can supply more examples (more will be supplied as we go on). The hypothesis I wish to put before you is this:

> Semantics and correspondence are co-extensive. *Whenever* two domains can be put into a correspondence (preferably, but not necessarily, a homomorphism), one of the domains (which can be considered to be the *syntactic domain*) can be understood in terms of the other (which will be the *semantic domain*).

1.4 COMPARISONS, PATTERNS, AND ROLES: A DIGRESSION.

To determine correspondences between two domains—a syntactic (or "new", not-yet-understood) domain and a semantic (or "given", antecedently-understood) domain—one makes *comparisons*. The result of a comparison is a determination that the "new" item "plays the same role" in *its* (syntactic) domain that the corresponding "given" item plays in *its* (semantic) domain. The two items are analogous to each other; a pattern seen in one domain has been matched or recognized in the other.

What are these "roles"? The *semantic* item's role is its *syntactic* role in the "given" domain. That is, *each* item—new *and* given—play roles in their respective domains. These roles are, in their respective domains, *syntactic* roles, that is, roles determined by relationships to other items in the domain. These relationships are not *cross*-domain relationships, but *intra*-domain relationships—that is, syntactic relationships, in Morris's sense.

But in what sense are these roles "the same"? They *correspond* to each other. But what does *that* mean? It means (1) that the two domains are both instances of a common pattern (which common pattern, as we just saw, is understood syntactically) and (2) that the new and given items both map to the same item in the common pattern. (For a detailed discussion of this general phenomenon, known as "unification", see Knight 1989b.) But then why not say that it's the common pattern that is the proper *semantic* domain, rather than say that the semantic domain is the "given" domain? Leo Apostel (1960: 2) suggested something like this: "If two theories are without contact with each other we can try to use the one as a model for the other or to introduce a common model interpreting both and thus relating both languages to each other." Typically, however, one uses as the "favored" semantic domain one that is "familiar". If one *did* take the common pattern as the semantic domain, the question of "same role" would arise again. But this time, there is no *other* common pattern, so there's no regress. But *now* what counts is the mapping between the two domains—the syntactic domain and either the "given" domain or the common pattern (it doesn't matter which). That mapping must have certain features, namely, the ones we identified above as characterizing semantic interpretation functions, such as being a homomorphism.

Again, what is the role of an item in the common pattern? That's a *syntactic* question. But before exploring that in more detail (in Chapters **??** and **??**), we need to look at semantic correspondences more carefully.

1.5 THE CORRESPONDENCE CONTINUUM: IMPLICATIONS.

There are three observations to be made about our data. First, *the syntactic domain need not be a "language"* in either the natural or formal sense. All that is required is that it be analyzable into parts (or symbols) that can be combined and related—in short, manipulated—according to rules. Apostel and Marx W. Wartofsky have made similar observations:

> Let then R(S,P,M,T) indicate the main variables of the modelling relationship. The subject S takes, in view of the purpose P, the entity M as a model for the prototype T. ... Model and prototype can belong to the same class of entities or to different classes of entities. The following possibilities immediately offer themselves: M and T are both images, or both perceptions, or both drawings, or both formalisms (calculi), or both languages, or both physical systems. All these possibilities have occurred. But we can also have the heterogeneous case: M can be an image, T a physical system, or inversely; M can be an image and T a perception; M can be a drawing and T a perception; M can be a calculus and T a theory or language; or inversely. M can be a language and T a physical or biological system. (Apostel 1960: 4.)

> The constraints of taking the model (or in the inverse logical image, the theory) as linguistic and the reference of the model (or the interpretation or embodiment of the theory) as extralinguistic ... seems unnecessarily restrictive. (Wartofsky 1966: 6.)

Second, *the so-called "syntactic" and "semantic" domains must be treated on a par*; that is, one cannot say of a domain that it is syntactic except relative to another domain which is taken to be the semantic one, and vice versa. Brian Cantwell Smith (1982b: 10) has made a similar observation:

> In a general sense of the term, *semantics* can be taken as the study of the relationship between entities or phenomena in a *syntactic domain S* and corresponding entities in a *semantic domain D* .... We call the function mapping elements from the first domain into elements of the second an **interpretation function** .... Note that the question of whether an element is syntactic or semantic is a function of the point of view; the syntactic domain for one interpretation function

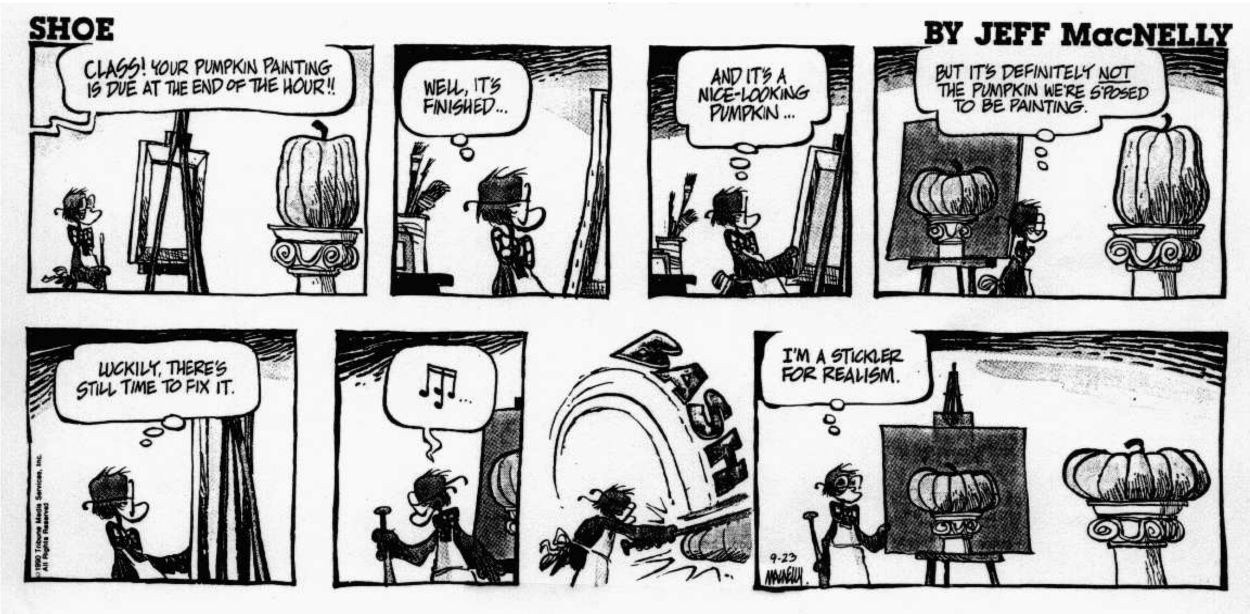

SHOE
BY JEFF MacNELLY
CLASS! YOUR PUMPKIN PAINTING IS DUE AT THE END OF THE HOUR!!
WELL, IT'S FINISHED...
AND IT'S A NICE-LOOKING PUMPKIN...
BUT IT'S DEFINITELY NOT THE PUMPKIN WE'RE S'POSED TO BE PAINTING.
LUCKILY, THERE'S STILL TIME TO FIX IT.
♫...
BASH
I'M A STICKLER FOR REALISM.
9-23
MacNELLY

Figure 1.2: How to make the semantic domain fit the syntactic domain.

> can readily be the semantic domain of another (and a semantic domain may of course include its own syntactic domain).

(I'll return to that closing parenthetical remark later (§1.7.3). Cf. the quotation from Apostel 1960: 4, above.)

Third, *what makes something an appropriate semantic domain is that it be antecedently understood*. This is, in fact, crucial for promoting semantics as "mere" correspondence to the more familiar notion of semantics as meaning or understanding. And, as indicated before, such antecedent understanding is, ultimately, syntactic manipulation of the items in the semantic domain.

Indeed, one can turn the tables. Suppose that something identified as the semantic domain is *not* antecedently understood, but that the putative syntactic domain *is*. Then by switching their roles, one can learn about the former semantic domain by means of its syntactic "interpretation". We saw one example of this in example 26 above. Another nice example of this for me was an article on "WHILE Loops and the Analogy of the Single Stroke Engine" (J. Cole 1991), in which the author uses the behavior of single-stroke engines to explain the behavior of while-loops. That works only to the extent that students antecedently understand single-stroke engines. I read the article conversely from how it was intended: I used my antecedent understanding of while-loops to help unravel the mysteries of the single-stroke engine! A similar point was made by Arturo Rosenblueth and Norbert Wiener (1945: 318), who gave the example of an "iron wire dipped in nitric acid" as a model of a nerve axon, pointing out that "the useful model in the pair" might really have been the "nerve axon instead of the wire".

In the worst case, if one knows *neither* domain antecedently, then one might be able to learn both together, in one of two ways: either by seeing the same structural patterns in both, or by "getting used to" them both. (Although, possibly, this contradicts the third observation, above.) In this case, neither is the syntactic domain—or else both are!

And if one wants to make the correspondences more exact (that is, to make the interpretation-function either total or onto), one can change *either* domain (as in the *Shoe* cartoon, Figure 1.2). Normally, one feels freer to change the syntactic domain, because that's the one that's treated as antecedently given, antecedently understood (hence the humor of the *Shoe* cartoon). That's what Bertrand Russell did in his analysis of definite descriptions (Russell 1905). But, as I have argued elsewhere, good arguments can be provided for changing the semantic domain (Rapaport 1981).

1.6 A HISTORY OF THE MUDDLE OF THE MODEL IN THE MIDDLE.

A number of people have made similar observations—that almost anything can be a model of almost anything else; that, therefore, there is no "privileged" state of being a model except, perhaps, that models must be antecedently understood; and that one person's syntactic domain might be another's semantic domain (the two-faced nature of models—the muddle of the model in the middle). In order to clarify these claims, as well as raise some other issues that will concern us later, let's look at what some of these people have had to say.

1.6.1 Rosenblueth and Wiener.

In their 1945 essay, "The Role of Models in Science", Rosenblueth and Wiener observe that scientists use models to understand the universe (p. 316). Thus, the universe (or, at least, data and observations) is the syntactic domain whose semantic interpretation is provided by a model, (part of) a scientific theory. In order to understand some part of the complex universe, one replaces it "by a model of similar but simpler structure" (p. 316)—this is the technique of *abstraction.* Note that, according to Rosenblueth and Wiener, one mark of being an abstraction is to be simpler than what it's an abstraction of; what it's an abstraction of (in this case, a part of the universe) will have "extra" features. These extra features might be quite important ones that are being ignored merely temporarily or for the sake of expediency, or they might be "noise"—irrelevant details. It is important to note that, for Rosenblueth and Wiener, the abstraction is the (semantic) model. Later, however, we will see that abstractions can also be seen as *syntactic* domains that can have *implementations* (Ch. 1). In such cases, the extra features not in the abstraction are often referred to as "implementation details".

There are, according to Rosenblueth and Wiener, two kinds of models: formal and material, both of which are abstractions (p. 316). Formal models seem to be more like "mathematical" models—that is, formal symbol systems, formal languages—in short, stereotypically syntactic domains. Material models, however, are not like stereotypical "semantic" domains; rather, they are more like scale models (p. 317).

"A material model is the representation of a complex system" (p. 317). This suggests that a material model represents some system that is itself material (for example, the solar system), not some mathematical/set-theoretic/linguistic/"syntactic" system (that's why it's not like a semantic interpretation in the sense of model theory). The material model "is *assumed* similar" (p. 317, my italics), although it can be "more elaborate" than that which it models (p. 318). This suggests that "implementation details"—that is, parts of the model that are *not* (or are not *intended* to be) representations of the complex system—are ignored. For instance, the physical matter that the model is made of, or imperfections in it, would be ignored: One does not infer from a plastic scale model of the solar system that the solar system is made of plastic.

"A formal model is a symbolic assertion in logical terms of an idealized relatively simple situation sharing the structural properties of the original factual system" (p. 317)—that is, a formal model is like a mathematical model—a syntactic system. One way to understand their claim is that there are three things: a formal model, an idealized situation, and a factual system. The formal model describes the idealized situation. But is it the formal model or is it the idealized situation that shares structural properties with the factual system? The answer, I think, is that it is the idealized situation. This seems to be literally what they say, and it corresponds closely to a claim of Smith's that we will examine later (§1.7.1).

However, there is another interpretation of the relationships among these three things, one that is, in a sense, a generalization of the first: Let the idealized situation be a *material* model of the factual system, and let the *formal* model express their shared structure:

> A material model may enable the carrying out of experiments under more favorable conditions than would be available in the original system. This translation presumes that there are reasonable grounds for supposing a similarity between the two situations; it thus presupposes the possession of an adequate formal model, with a structure similar to that of the two material systems. (p. 317.)

Why is a formal model "presupposed"? Note that the formal model would have to model *both* the material model *and* the original system, much like the notion of a common pattern that we discussed earlier (§1.4). Here is a possible explanation of the presupposition: Let O be the original system. Let $M_m(O)$ be a material model of O. To be able to use $M_m(O)$ for scientific purposes, one wants to be able to argue that if $M_m(O)$ has some property P, then so does O (or, perhaps, that if $M_m(O)$ has some property P_{M_m}, then O has the property P, where $P_{M_m} = M_m(P)$). But to do this, one needs a *theory* that says that $M_m(O)$ and O are relevantly structurally alike. That theory would be a formal model M_f that would be simultaneously a model of O and of $M_m(O)$; that is, it would be such that $M_f(O) = M_f(M_m(O))$—it would "embody" (if you will excuse that rather metaphorical expression!) the common structure of O and $M_m(O)$.

How does this help in understanding O? "Material models ... may assist the scientist by replacing a phenomenon in an unfamiliar field by one in a field in which he [sic] is more at home" (p. 317). That is, the material model is antecedently understood. Rosenblueth and Wiener observe that in the 18th and 19th centuries, mechanical models were used to understand electrical problems, but that in the 20th century, electrical models were used to understand mechanical problems! One person's antecedently understood domain is another's in need of understanding. A formal model can "suggest a material one" (p. 318). That is, the abstract formal model can be "embodied", the "converse" of abstraction (p. 320); it is what I have called "implementation". Thus, one begins with O; one can then construct $M_f(O)$, and use this to develop $M_m(M_f(O))$, which will be an $M_m(O)$. But "[t]he formal model need not be thoroughly comprehended; the material model then serves to supplement the formal one" (p. 317). If $M_f(O)$ is not antecedently understood, then $M_m(O)$ can be used to understand it—the material model of O can be used to understand the formal model of O. Better yet—and consistent with my hypothesis—each can be used to (help) understand the other: The abstract formal model can be constructed to help us to understand the original system as well as a material model of the original system, or the abstract formal model can be implemented to produce a material model of it, which can then be used to understand the original system.

What happens if the model is precisely as complex as the original?

> ... it will become that system itself. That is, in a specific example, the best material model for a cat is another, or preferably the same cat. In other words, should a material model thoroughly realize its purpose, the original situation could be grasped in its entirety and a model would be unnecessary. (Rosenblueth & Wiener: 320).[5]

Of course, there is a difference between O itself (say, a cat) and *another* system O' (say, another cat) that serves as $M_m(O)$. Granted, if O itself can be understood in and by itself, then no $M_m(O)$ would be needed, although sometimes one must use an equivalent but *distinct* O' (for reasons, say, of convenience). (How does one study O in and by itself? By getting used to it—that is, syntactically! But, again, I anticipate myself.) One can study the behavior and biological properties of cats in general (and of my cat in particular, at least insofar as it is representative of cats in general) by studying the behavior and biological properties of *your* cat. One then argues by analogy: if O' has property P, then (in all likelihood), O has P.

But it does not follow that models are unnecessary. In fact, they are unavoidable: Granted, the best way to study cats in general is to study a particular real cat rather than a model of a cat. And the best way

[5] They cite Lewis Carroll's *Sylvie and Bruno* on a map that is the country itself. Cf. Josiah Royce's *The World and the Individual*, Vol. 1 (1899), cited in Borges 1981: 234, Rapaport 1978: 164, and Eco 1982.

to study a particular cat is to study *it*, not some other cat (although controls are useful). But the inevitable result of such a study is a model or theory of that cat (or of cats in general)!

Of course, with the exception of those inquiries in which a specific *O* is used as a representative sample of *O*s in general (that is, as a model of itself), models that are as complex as that which they are designed to help us understand are unlikely to be of much use. This is one of the difficulties with many connectionist models of cognition: Their complexity approaches that of the cognitive behavior they are intended to model (or, to reproduce), and they do not seem to have any features that explain their behavior. Certain inputs are provided, certain weights are adjusted according to algorithms that are independent of the cognitive behavior being modeled, and—lo and behold—appropriate outputs appear. But what do the various weights and adjustments mean with respect to the particular cognitive behavior? If we don't understand the connectionist system, it doesn't really *tell* us anything about cognition. (As Joseph Weizenbaum (1976: 40–41), observed, "Indeed, we are often quite distressed when a repairman returns a machine to us with the words, 'I don't know what was wrong with it. I just jiggled it, and now it's working fine.' He [sic] has confessed that he failed to come to understand the law of the broken machine and we infer that he cannot now know, and neither can we or anyone, the law of the 'repaired' machine. If we depend on that machine, we have become servants of a law we cannot know, hence of a capricious law. And that is the source of our distress.") In other words, for something to be used as a model of another thing, it must be antecedently understood.

Rosenblueth and Wiener conclude by arguing that partial models are all we can ever get, because our minds are finite. What is the implication of this for computational cognitive science? Computational cognitive scientists (try to) create a (partial) model of cognition by means of an algorithm that can then be implemented in a computer. Can we ever get the *full story* of cognition this way? Possibly: Though we might not understand a "complete 'model'" (that is, a self-model) *directly*, we might be able to understand it by successive approximation. We can fully understand a partial model, and then augment it by a small, understandable amount. In fact, though, this would be fraught with all the problems that one faces when small changes are introduced into software: One small change *here* might have untold effects *there*, where "there" might be several thousands of lines of code away. However, for the case of cognition, it might well turn out that there is a threshold beyond which it's unnecessary to go in order to have created a cognitive agent.

(Their essay is interesting for two other reasons. First, they distinguish between "closed-box" and "open-box" problems (pp. 318–319). This is surely an early version of the notions of "black boxes" and "glass boxes". Second, they base an early version of homuncular functionalism on this: "Scientific progress consists in a progressive opening of these [closed] boxes" and subdividing closed boxes into "several smaller shut compartments" some of which "may be ... left closed, because they are considered only functionally, but not structurally important" (p. 319).)

1.6.2 Wartofsky and the Model Muddle.

> All this is by way of arguing for a representationalist account of models. But 'representation' then is taken in the broadest sense as any sort of mapping of structures on structures, or qualities on qualities. The essential feature of representation is reference, and it may be argued that not all reference is 'representational'. I would argue, perhaps perversely, that it is. (Wartofsky 1966: 8.)

I owe the phrase 'the model muddle' to Marx W. Wartofsky's 1966 essay of that name:

> The symptom of the muddle is the proliferation of strange and unrelated entities which come to be called models. Thus 'model' is used for the straightforward mechanical model ...; as well, for the theoretical construct in physics or in psychology which has its embodiment only in mathematical or verbal inscriptions or utterances ...; and equally, for the mapping of some uninterpreted formal system on some interpretation or embodiment of it .... (p. 1.)

This is the proliferation I exhibited in §1.3. Wartofsky's move is to classify all of these notions "as species of the genus representation; and to take representation in the most direct sense of image or copy" (p. 1). In a later essay, which we will turn to shortly (Wartofsky 1979), he takes "representation" in the sense of "reminder", which I think is slightly more general than "copy" or "image", though not quite as general as "correspondence".

What I called "the muddle of the model *in the middle*" is expressed by Wartofsky as follows:

> Inverse to the ordinary view of models as abstractive representations of some object or state of affairs, logicians speak of models as the interpretations or embodiments of some formal calculus, in which the relation of isomorphism (more strictly, homomorphism) holds between the structure of the formal system and that of its interpretation. (p. 4.)

The way to resolve the muddle is to put the model in the middle, thus:

$$\text{formal system} \rightarrow \text{model} \rightarrow \text{actual world (objects, states of affairs)}$$

The model abstractively represents (aspects of) the actual world. It also is an "interpretation or embodiment"—an *implementation*—of a formal system. But the formal system also abstractively represents the actual world—and with a vengeance, since the model of the formal system will typically have "implementation details", just as the actual world has "implementation details" with respect to—that is, is more complex than—the model. And, in this case, the formal system abstractively represents the *model*, too.

Wartofsky offers a number of theses about his general notion of model. Let us take a look at them.

1. One of Wartofsky's fundamental assumptions is "that between any two things in the universe there is some property they both share, there is some relation which they bear to each other" (p. 4):

   $$\forall xy \exists \mathrm{P}[\mathrm{P}x \wedge \mathrm{P}y].$$

   (What is the "*relation* which they bear to each other"? Presumably, it is the relation of sharing a common property.) Is this plausible? How is a raven like a writing desk? Well, they are both physical objects. How is a physical object like an abstract object (how is the Eiffel Tower like the set of all unicorns)? Well, they are both capable of being objects of thought (or, in this case, they are both used as examples in this section!). So, perhaps with a bit of stretching, one *can* find a common property for any two things. In most ordinary cases, though, one probably won't have to stretch too far (this is what makes metaphors so common). And, as Wartofsky later notes (1979: xx), citing Nelson Goodman, "everything has infinitely many properties in common with everything else".

2. The modeling relation is triadic (p. 6):

   $M(S, x, y)$ means: cognitive agent S takes x as a model of y.

   The crucial point here is that modeling is not an objective or mind-independent relation between two entities. Rather, it is relative to a cognitive agent—to "cognitive activity" (p. 4).

3. Given (1), the modeling relation can be defined as (or in terms of) representation (p. 4):

   Let S be a cognitive agent.
   Let x, y be two entities.
   Let P be one of their common properties, as guaranteed by (1).
   Let P_x be x's instantiation of P (and similarly for y).
   Then $M(S, x, y) =_{df} S$ takes P_x as representing P_y.

4. Wartofsky posits "a trivial truth: models exist" (p. 3):

$$\exists SxyM(S,x,y).$$

That is, there are things x, y such that x reminds S of y because of properties they share.

5. "[A]n additional trivial truth ...: anything can be a model of anything else! This is to say no more than" (1), above: $\forall xy \exists \mathrm{P}[\mathrm{P}x \wedge \mathrm{P}y]$ (p. 4). However, it says something rather different, since the "trivial truth" is modal, whereas (1) is not. The "trivial truth" seems to be this:

$$\forall Sxy \diamond M(S,x,y).$$

The idea seems to be that *because* any two things have a common property, anyone *could* take one as a model of the other. (Cf. Wartofsky 1979: xx.)

6. Nevertheless, "there are clearly only some things which we choose to sort out as models of some other things ...." (p. 4): The force of 'only' suggests the following interpretation:

$$\exists Sxy \neg M(S,x,y).$$

That is, there are some things that no one takes as models of other things.

7. However, there is "a simple constraint on models, which we may take as a definition (or part of one), or as a convention: nothing which is a model is to be taken as a model of itself, nor of something identical with it" (p. 4):

$$\forall Sx \neg M(S,x,x).$$

Wartofsky observes that "In a weak sense, one may enforce the constraint by stating that at the limit, the case of anything being a model of itself is trivial. But Rosenblueth and Wiener are willing to go all the way ..." (p. 5).

8. Under this constraint, M is asymmetrical (p. 5). Yet Wartofsky rejects the following natural interpretation of the asymmetry:

$$\neg(M(S,x,y) \rightarrow M(S,y,x))$$

on the grounds that it is not merely that the entities x and y cannot be switched, but rather that in order for S to take x as a model of y, x must (be believed by S to) have fewer relevant properties than y (pp. 5–6): A model "has to be less rich in the range of relevant properties than its object", because if it were "equally rich in the same properties ... it would be identical with its object", and if it were "*richer* in properties, ... these would then not be ones relevant to its object; it [the object] wouldn't possess them, and so the model couldn't be taken to represent them in any way" (pp. 6–7).

But I think it is more appropriate to locate the asymmetry in the fact that the model must be antecedently understood: Suppose that M is an antecedently understood model of some state of affairs or object O. Suppose, first, that M has fewer properties than O, the case that Wartofsky takes to be the norm. Here, the asymmetry between M and O could be ascribed either to M's having fewer properties (as Wartofsky would have it) or to M's being antecedently understood (as I would have it), so we cannot distinguish between our two positions on these grounds. Suppose, next, that M and O have the same properties. On Wartofsky's view, the asymmetry is lost, but if I antecedently understood M, I can still use M as a model of O: This is the Rosenblueth and Wiener cat-case. It is also the situation Daniel C. Dennett describes in his Ballad of Shakey's Pizza Parlor (Dennett 1982: 53–60): Since all Shakey Pizza Parlors are indistinguishable, I can use my knowledge of one of them to help me understand the others (for example, to locate the rest rooms). Similarly, I know how *your* ball-point pen works, because it's just like mine. Finally, suppose that M has *more* (or perhaps merely *different*) properties than O. For example, one could use (the liquidity of) milk as a model of (the liquidity of) mercury (at least, for certain purposes, though not for understanding its meniscus),[6] even though milk has more (certainly, different) properties. These extra (or different) properties are precisely what I have called "implementation details"; but they are *merely* that—hence, to be ignored. As long as I antecedently understand M, I can use it as a model of O, no matter how many properties it has. But if I *don't* antecedently understand M, then I *can't* use it as a model (except in the very special case, mentioned earlier, in which I lack antecedent understanding of *both* M *and* O, and use them together to understand them both).

Nevertheless, the crucial feature of Wartofsky's theory is thesis (3), his definition of models as representations, for it is in virtue of this that we can see why anything can be a model of anything else (except possibly itself) and hence why it is that one person's syntactic domain can be another's semantic one (and vice versa): I might take x (or P_x) as representing y (or P_y), whereas you take y as representing x. But we can go one step further than Wartofsky: The reason why I take x as representing y (rather than the other way round, as you do) is that I am more familiar with x, I antecedently understand it. And how do I do that? Why is it that I understand x? Because I am used to it.

1.7 THE CORRESPONDENCE CONTINUUM OF BRIAN CANTWELL SMITH.

What I have referred to as the "correspondence continuum" and the "muddled models" of Rosenblueth, Wiener, and Wartofsky has received its most explicit statement and detailed investigation in the writings of Brian Cantwell Smith (from whom I have borrowed the term 'correspondence continuum').

1.7.1 Preliminary Observations: Worlds, Models, and Representations.

In an important essay on computer ethics, "Limits of Correctness in Computers" (1985), Smith sets up the "model muddle" as follows:

> When you design and build a computer system, you first formulate a model of the problem you want it to solve, and then construct the computer program in its terms. ...
>
> To build a model is to conceive of the world in a certain delimited way. ... computers have a special dependence on these models: *you write an explicit description of the model down inside the computer*, in the form of a set of rules or what are called *representations*—essentially linguistic formulae encoding, in the terms of the model, the facts and data thought to be relevant to the system's behaviour. ... In fact that's really what computers are (and how they differ from other machines): they run by manipulating representations, and representations are always formulated in terms of models. This can all be summarized in a slogan: no computation without representation. (p. 636.)

The picture we get from this (incorporating some additions to be discussed shortly) is shown in Figure 1.3 (cf. Smith 1985: 639): The model, M, is an abstraction, R_1, of the real-world situation W (Smith 1985: 637): It is the world conceived "in a certain delimited way." For instance, "a hospital blueprint would pay attention to the structure and connection of its beams, but not to the arrangements of proteins in the wood the beams are made of ..." (Smith 1985: 637). The model M is itself "modeled", or *described* (R_2), in the computer program P; the model, thus, is a "swing domain", playing the role of syntactic domain to the real world's semantic domain, and the role of semantic domain to the computer program's syntactic—indeed, linguistic—description of it (cf. Smith 1985: 637).

Smith calls the process of abstraction (which for him includes "every act of conceptualization, analysis, categorization", in addition to the mere omission of certain details) a necessary

[6] This milk/mercury example is due to V. Kripasundar.

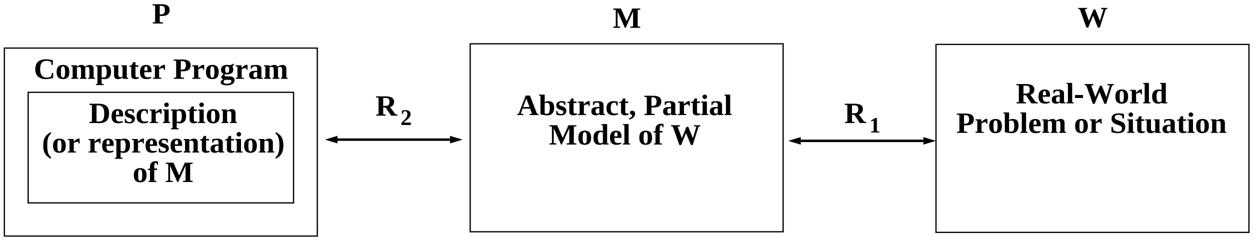


Figure 1.3: Smith's version of the "model muddle".

> act of violence—[if you] don't ignore some of what's going on—you would become so hypersensitive and so overcome with complexity that you would be unable to act. (Smith 1985: 637.)

Of course, one ought to do the least amount of violence consistent with not being overwhelmed. This might require successive approximations to a good model that balances abstraction against adequacy. George Lakoff's complaints about what he calls "objectivism" (in *Women, Fire, and Dangerous Things* (Lakoff 1987))—specifically, his objections to "classical" categories defined by necessary and sufficient conditions—can be seen as a claim that "classical" categories do too much violence, so that the resulting models are inadequate to the real-world situations. What's needed are better approximations to—better models of—reality (which, for Lakoff, take the form of "idealized cognitive models").

In view of our discussion of Rosenblueth and Wiener's self-modeling cat, we may ask of Smith why complexity makes acting difficult. Doesn't the real-world situation have precisely the maximal degree of complexity? Yet a human—a real-world cognitive agent to be modeled by the techniques of AI—is capable of acting. Moreover, a complete and complex model of some real-world situation might be so complex that a mere human trying to understand *it* might "drown" in its "infinite richness" (Smith 1985: 637), much as a human can't typically hand-trace a very long and complex computer program. Yet a computer can execute that program without "drowning" in its complexity.

But for Smith,

> models are inherently *partial*. All thinking, and all computation, are similarly partial. Furthermore—and this is the important point—thinking and computation *have* to be partial: that's how they are able to work. (Smith 1985: 637.)

Note that some of the partiality of thinking and computation is inherited from the partiality of the model and is then compounded: To the extent that thinking and computation use partial descriptions of partial models of the world, they are doubly partial. Much inevitably gets lost in translation, so to speak. Models certainly need to be partial at least to the extent that the omitted details (the "implementation details") are irrelevant and certainly to the extent that they (or their descriptions) are discrete whereas the world is continuous. But does thinking "have to be partial" in order to be "able to work"? A *real* thinking thing isn't partial—it is, after all, part of the real world—though its descriptions of models of the world might be partial. And that's really Smith's point—thinking things (and computing things) work with partial models. They "represent the world *as being a certain way*" (Smith 1987: 51n1), "*as being one way as opposed to another*" (Smith 1987: 4): They present a fragmentary point of view, a facet of a complete, complex real-world situation—they are objects under a (partial) description (cf. Castañeda's (1972) "guises"; see §§**??**, **??**, below).

So we have the following situation. On one side is the real world in all its fullness and complexity. On the other side are partial models of the world and—embedded in computer programs—partial descriptions of the models. But there is a gap between full reality, on the one hand, and partial models and descriptions, on the other, insofar as the latter fail to capture the richness of the former, which they are intended to interact with: Action "is not partial .... When you reach out your hand and grasp a plow, it is the real field you are digging up, not your model of it ... [C]omputers, like us, participate in the real world: they take real actions" (Smith 1985: 637–638). This holds for natural-language competency programs, too. Their actions are speech acts, and they affect the "full-blooded world" (Smith 1985: 637) to the extent that communication between them and other natural-language–using agents is successful.

To see how the "reaching out" can fail to cross the gap, consider a blocks-world robot I once saw. It was a simple device that could pick up and put down small objects at various locations in an area that was about one yard square. It had been programmed with a version of an AI program for doing such blocks-world manipulations that appears in Patrick H. Winston's (1975) AI text. Now this robot really dealt with the actual world—it was not a simulation. But it did so successfully only by accident. If the blocks were *perfectly* arranged in the blocks-world area, all went well. But if they were slightly out of place—as they were on the day I saw the demo—the robot would blindly and blithely execute its program and behave as if it were picking up, moving, and putting down the blocks. More often, it failed to pick them up, knocked them down as it rotated, and dropped them if it hadn't quite grasped them at the right angle. It was really quite humorous, if not downright pathetic, to watch. The robot was doing what it was "supposed" to do, what it was programmed to do, but its partial model was inadequate. Its *successful* runs were, thus, accidental—they worked only if the real world was properly aligned to allow the robot to affect it in the "intended" manner. (Smith 1985: 637–638 describes a similar example). Clearly, a robot with a more complete model would do better. The checkers-playing robot at the University of Rochester, for example, has a binocular vision system that enables it to "see" what it's doing and to bring its motions into alignment with a changing world (Marsh, Brown, LeBlanc, Scott, Becker, Das et al. 1992; Marsh, Brown, LeBlanc, Scott, Becker, Quiroz et al. 1992).

A theme that will become more important later on begins to emerge. Computers participate in the real world *without interpretations of their behavior by humans* and without the willing participation of humans. (Although I will be concerned here only with the implications of this for computational cognitive science, it is important to see that there are *moral* implications, too, which are the ones Smith emphasizes in his essay.) Consider a program with natural-language competence. Does it really "use language" or "communicate" without a human interpreter? There are two answers: 'yes' and 'no, but so what?'. Let me briefly present these now; I'll say more about them as we go on.

Yes. As long as the natural-language–using computer is using the vocabulary of some natural language according to the rules of grammar of that language, it is thereby using that language, even if there is no other language-using entity around, including a human. This is true for humans, too: As Kah-Kyung Cho has observed, even if I talk to myself without uttering a sound, I mean things by my silent use of language. Sound or other external signs of language-use are not essential to language.[7] And, therefore, neither is a hearer or other interlocutor (who is distinct, extensionally speaking (cf. Shapiro 1986), from the speaker). (Though without an interlocutor, it could not pass the Turing Test; cf. §??.)

No; but so what? A human might interpret the computer's natural-language output differently from how the computer "intended" it. Or one might prefer to say that the computer's output is meaningless until a human interprets it. The output would be mere syntax; its semantics would have to be provided by the human, *although it could be provided by another natural-language–using computer*. However, the same situation can arise in human-to-human communication. Nicolaas de Bruijn once told me roughly the following anecdote: Some chemists were talking about a certain molecular structure, expressing some difficulty in understanding it. De Bruijn, overhearing them, thought they were talking about mathematical lattice theory, since everything they said could be—and was—interpreted by him as being about the mathematical domain rather than the chemical domain. He knew the solution of their problem in terms of lattice theory, and told it to them. They, of course, understood it in terms of chemistry. Were de Bruijn and the chemists talking about the same thing? No; but so what? They *were* communicating!

[7] I owe this point to Cho's lecture, "Rethinking Intentionality," SUNY Buffalo Center for Cognitive Science, 7 November 1990. Cf. Cho 1992.

It is also important to note that when a natural-language–competent computer interacts with a human or another natural-language–competent computer, both need to be able to reach a more-or-less stable state of mutual comprehension. If the computer uses an expression in an odd way (perhaps merely because it was poorly programmed or did not adequately learn how to use that expression), the human must be able to correct the computer—*not* by reprogramming it—but by *telling* it, in natural language, what it should have said. Similarly, if the human uses an expression in a way that the computer does not recognize, the computer must be able to figure out what the human meant. These are issues I have dealt with before (Rapaport 1988), and will deal with again, below (§1.8.2, and Chs. **??** and **??**).

1.7.2 The Model–World Gap and the Third-Person Point of View.

The gap between model and world is difficult, perhaps impossible, to bridge:

> . . . we in general have no guarantee that the models are right—indeed we have no *guarantee* about much of anything about the relationship between model and world. . . .
>
> In philosophy and logic . . . there is a very precise mathematical theory called "model theory." You might think that it would be a theory about what models are, what they are good for, how they correspond to the worlds they are models of . . . . Unfortunately, . . . model theory doesn't address the model–world relationship at all. Rather, what model theory does is to tell you how your descriptions, representations, and programs *correspond to your model.* (Smith 1985: 638.)

To "address the model–world relationship" requires a language capable of dealing with *both* the model *and* the world. This would, at best, be a "Russellian" language that allowed sentences or propositions to be constructed out of real-world objects (Russell 1903, Moore 1989).[8] It would have to have sentences that explicitly and directly linked parts of the model with parts of the world (reminiscent, perhaps, of the way that Helen Keller at the well house was herself the link between the world—with water running over one hand—and her language—with 'w-a-t-e-r' simultaneously being finger-spelled into the other). But how can such model–world links be made? The only way, short of a Russellian language, is by having *another* language that describes the world, and then provide links between *that* language and the model. (In fact, that would have to be done in a meta-language. I am also assuming, here, that the model is a language—a description of the world. If it is a non-linguistic model, we would need, then, yet another language to describe *it.*) But this leads to a regress with a Zenoesque or Bradleyesque flavor, for how, then, will we be able to address the relationship between the world and the language that describes it? This parallels the case of the mind, which, insofar as it has no direct access to the external world, has no access to the reference relation.

Model theory, as Smith points out, discusses only the relation between a model and its description—relation R_2 in Figure 1.3. It does not deal with relation R_1. Two questions need to be answered: *Could* it discuss R_1? *Does* it deal with R_2? By my hypothesis that semantics is correspondence, the two cases should be parallel; one ought to be able to deal with both R_1 and R_2, or with neither. But we have just seen that R_1 cannot be dealt with except indirectly. Consider R_2. Is it the case that the relation between the computer and the model is dealt with by model theory? No; as Smith says, it deals with the relation between a *description* of the model and the model. After all, the computer is part of the real world (cf. Rapaport 1985/1986: 68, Fig. 1). So the argument about the model–world relationship also holds here, for, in the actual computer, there is a physical (real-world) implementation of the model.

How, then, can a relation between a syntactic domain and a semantic domain be understood? Only by taking an independent, external, third-person point of view. There must be a standpoint—a language, if you will—capable of having equal access to *both* domains. A semantic relation can obtain between two domains, but neither domain can describe that relation by itself. From the point of view of the model, nothing can be said about the world. Only from the point of view of some agent or system capable of taking

[8] Cf. Helen Keller's labels; see Ch. **??**.

both points of view simultaneously can comparisons be made and correspondences established. This, too, will loom larger in what follows.

Here is another way to approach this. Smith offers a Kantian analogy:

> Mediating between ... ["a description, program, computer system (or even a thought—they are all similar in this regard) ... and the very real world"] is the ... model, serving as an idealized or preconceptualized simulacrum of the world, in terms of which the description or program or whatever can be understood. One way to understand the model is as the glasses through which the program or computer looks at the world: it is the world, that is, as the system sees it (though not, of course, as it necessarily is). (Smith 1985: 638.)

If the model is placed in the role of the external observer with access to both the computer program's indirect description of the world and the world itself, still—from the point of view of the computer—the computer has no direct access to the world. Similarly for human use of natural language: A hearer must construct a mental model *of* the *speaker's* model of the world, but cannot have direct access to the speaker's model. We can only deal with Kantian phenomena, not with Kantian noumena (cf. Castañeda 1989c: 35).

1.7.3 The Continuum.

Smith sees the classical semantic enterprise as a special case of a general theory of correspondence. I see *all* cases of correspondence as being semantic. Perhaps this is little more than a terminological difference, since we both emphasize correspondence.

Smith begins his 1987 essay "The Correspondence Continuum" by considering such core semantic or intentional relations as representation and knowledge, "asymmetric" relations (that is, ones such that $\neg(xRy \rightarrow yRx)$ that "characterise phenomena that are *about* something, that refer to the world, that have meaning or content" (Smith 1987: 2). As we've seen, given two domains x and y, either can be used to represent the other, possibly even at the same time. Insofar as there is an asymmetry, it is to be located in one domain's being antecedently understood, as I argued above (§1.6.2).

In an earlier, influential, essay, "Reflection and Semantics in a Procedural Language" (1982), Smith enunciated his Knowledge Representation Hypothesis:

> Any mechanically embodied intelligent process will be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantical attribution, play a formal but causal and essential role in engendering the behaviour that manifests that knowledge. (Smith 1982: 33.)

In "The Correspondence Continuum", he elaborates on this by presenting a "two-factor" analysis of knowledge representation (Smith 1987: 3). On this analysis, there is an agent with internal, contentful, and causal structures, which he calls 'impressions'. These are contrasted with 'expressions', which are "elements of an external language" (Smith 1987: 3). For concreteness, think of the nodes in Cassie's mind, or the terms of a language of thought, as impressions.

The first factor of the two-factor analysis of knowledge representation is that impressions have a "(functional) role" (Smith 1987: 4). That is, they are causally produced by the agent's previous behavior and experiences, they play a role in causing the agent's future behavior, and they are "manipulable", that is, they can be combined to produce more impressions. This is essentially Wilfrid Sellars's theory of the language game (1955/1963; cf. §??, below). Thus, the domain of impressions is a syntactic domain (impressions are manipulable).

The second factor is that impressions have "(representational) import" (Smith 1987: 4). That is, there is a content relation R such that if $a\mathrm{R}b$, then a is an impression and b is a state of affairs in the world that includes the agent. 'Import' is a nice term: Representations import fragments of the external world into the mind. As we saw (§1.7.1), R is (typically) partial; impressions represent *aspects* of the world. Thus, the domain of impressions is not only a syntactic domain, but also a semantic one.

The two factors are merged in "the *full significance* of an impression" (Smith 1987: 5). Presumably, this is a close cousin of (if not identical to—cf. Smith 1987: 53n14) his earlier "general significance function ... that recursively specifies ... together" the syntactic relations among impressions and the "designation relation"—the import—between impressions and the world (Smith 1982; cf. §1.3, example 27). The two-factor analysis is the Knowledge Representation Hypothesis (Smith 1987: 5). Smith notes that the two factors need not be independent and that functional role need not "arise solely from *syntactic* properties of the representational structures" (Smith 1987: 5–6), though it is not clear what he means by 'syntactic' here.

He gives a very general characterization of the semantic enterprise as taking a "source" domain (the syntactic domain—for example, a set of impressions in a knowledge representation system), a semantic domain (a "target" domain), and an *extensional* interpretation function from the source syntactic domain to the target semantic domain (p. 8). This suggests that compositionality is *not* an essential constraint on semantics—that, in fact, there are no constraints at all. Indeed, he observes that this does not distinguish the semantic relation from an arbitrary one. However, there are different varieties of semantic relations, depending on further conditions:

> But in practice more assumptions are adopted. I will label as *model-theoretic* those semantical analyses that accept (which I don't!) the following additional claims:
>
> 1. The elements of the representational domain are assumed to be *linguistic*. ... [that is,] linear sequences of some sort ... with an inductively specified recursive structure founded in an initial base set of atomic elements called a *vocabulary*, and assembled according to rules of composition specified in a *grammar*. Furthermore, the interpretation relation is usually defined *compositionally*, so that its meanings (not contents!) are assigned both to the vocabulary items and to the recursive structures engendered by the grammatical rules, in such a way that the meaning of a complex whole arises in a systematic way from the meaning of its parts. (Smith 1987: 8–9.)

I take it that by 'representational domain' he means the source syntactic domain. I agree that it is not necessary for the syntactic domain to be thus linguistic—consider the variety of syntactic domains we saw in §1.3. Note, too, that being linguistic is *not* a restriction on the target semantic domain, yet it would have to be for "swing" domains if one accepted the model-theoretic view.

By 'meanings' vs. 'contents', Smith is alluding to the distinction between "meaning" as "what all instances or uses of a given structure type have in common" and "content" or "interpretation" as "what a particular use or instance of that type refers to" (Smith 1987: 7). For example, the *meaning* of the first-person pronoun 'I' is a projection *function* that takes a speaker–time–location triple and returns the speaker, whereas the *content* of 'I' for a specific speaker S_0 at a specific time T_0 at a specific location L_0 *is* S_0, the *speaker* him- or herself (and not a function).

Compositionality presumably only makes sense for "linguistic" syntactic domains. Smith goes on (p. 8) to indicate that there are degrees of compositionality, ranging from "strong" (in which the meaning of a whole is a function of the meanings of its parts) to "weak" (in which the meaning of a whole is "constrained" by "systems of regularities among the parts"—which might, for example, account for idioms or interjections (on the latter, cf. Wilkins 1992, 1995)).

The second assumption of model-theoretic semantics is this:

> 2. In a case where the elements of syntactic domain S correspond to elements of semantic domain D_1, and the elements of D_1 are themselves linguistic, bearing their own interpretation relation to another semantic domain D_2, then the elements of the original domain S are called *metalinguistic*. Furthermore, the semantic relation is taken to be *non-transitive*,

> thereby embodying the idea of a strict use–mention distinction, and engendering the familiar hierarchy of metalanguages. (Smith 1987: 9.)

However, in the case Smith has in mind, it's not clear that S really *is* linguistic (although D_1 *is*), for S will typically consist of *names* of items in D_1, but names are not linguistic in the sense of the first assumption above. Second, suppose that S = French, D_1 = English, and D_2 = the actual world. Then the semantic relation *is* transitive, and there is *no* use–mention issue. Here, I am thinking of a machine-translation system, *not* of the case of a French-language textbook written in English (that is, a textbook whose object language is French and whose metalanguage is English). Clearly, though, there *are* systems of the sort described in this assumption.

There are two more assumptions:

> 3. ... whatever information disambiguates a given use of an otherwise ambiguous expression is included as a parameter of meaning; content is then obtained from the meaning by fixing that parameter. ... Thus ... dependence on circumstantial or contextual factors [is] folded into the interpretation. (Smith 1987: 9–10.)
> 4. It is not necessary ... that the semantic domain be the real domain that the expressions are about. Rather, the semantic domain is required to be a set-theoretic structure, viewed as a *model* of the real semantic domain. (Smith 1987: 10.)

Assumption 3 seems to be that the interpretation function maps elements of the syntactic domain paired with circumstantial parameters to elements of the semantic domain. Since the circumstantial parameters are presumably part of the semantic domain, this might explain why Smith says that his two factors are not independent. Assumption 4, of course, is the model–world gap.

Smith clarifies and modifies the picture presented in Assumption 3 by pointing out, in connection with Assumption 4, that there is a "modeling relation" between the semantic domain and the actual world as well as a "genuine interpretation function" from the syntactic domain paired with circumstantial parameters to the actual world. Why is one a *relation* and the other a *function*? In any case, his point is the now-familiar one that in model-theoretic semantics, the modeling of the actual world, which produces a set-theoretic semantic domain, is not normally paid attention to; it is "free" or "invisible" (p. 10). Presumably, the diagram commutes: The composition of (1) the model-theoretic interpretation function from syntactic-domain–plus–circumstantial-parameters to semantic domain with (2) the modeling relation between the semantic domain and the actual world yields the same results as the genuine interpretation function (see Figure 1.4).

A further point, and this is where the notion of a correspondence continuum first seems to appear, is that there are "complex situations that include both modeling and iterated representation of the sort discussed in the second assumption" (p. 11). The picture we have is shown in Figure 1.5. To see an example of this in detail, consider Smith's discussion of programs and processes, where programs are "inert linguistic entities, built up of *expressions*; processes, in contrast, are active, manifest behaviour, and are comprised of *impressions*" (Smith 1987: 17; my italics). The process is part of the actual world; it thus has to be modeled to be dealt with (by Assumption 4). We have, then, in Figure 1.6, a version of Figure 1.4 (cf. Smith 1987: 18, Fig. 7). Both relations here are semantic: "modelling ... is itself a semantic, intentional, notion" (Smith 1987: 23); that is, the relation between the actual world (C) and a set-theoretical model of it (M_C) is semantic, and the set-theoretical model (M_C) is in turn the semantic domain for model-theoretic semantics (P). But this is only part of the story, since process C is, after all, the dynamic result of program P's static modeling of some part of the actual world, and the actual world (W) can, independently of P, be set-theoretically modeled (say, by M_W), as in Figure 1.7 (cf. Smith 1987: 18, Fig. 8). Smith says that "one is apt to identify ... M_C ... with ... M_W" and that W is what C "is genuinely about" (p. 18), but it seems to me that we don't have to worry about non-transitive use–mention problems here: C *is* a model of W. (And, of course, it is part of W, as is everything.)

There is more: For one thing, the process, C, is, typically, implemented on "a lower-level machine".

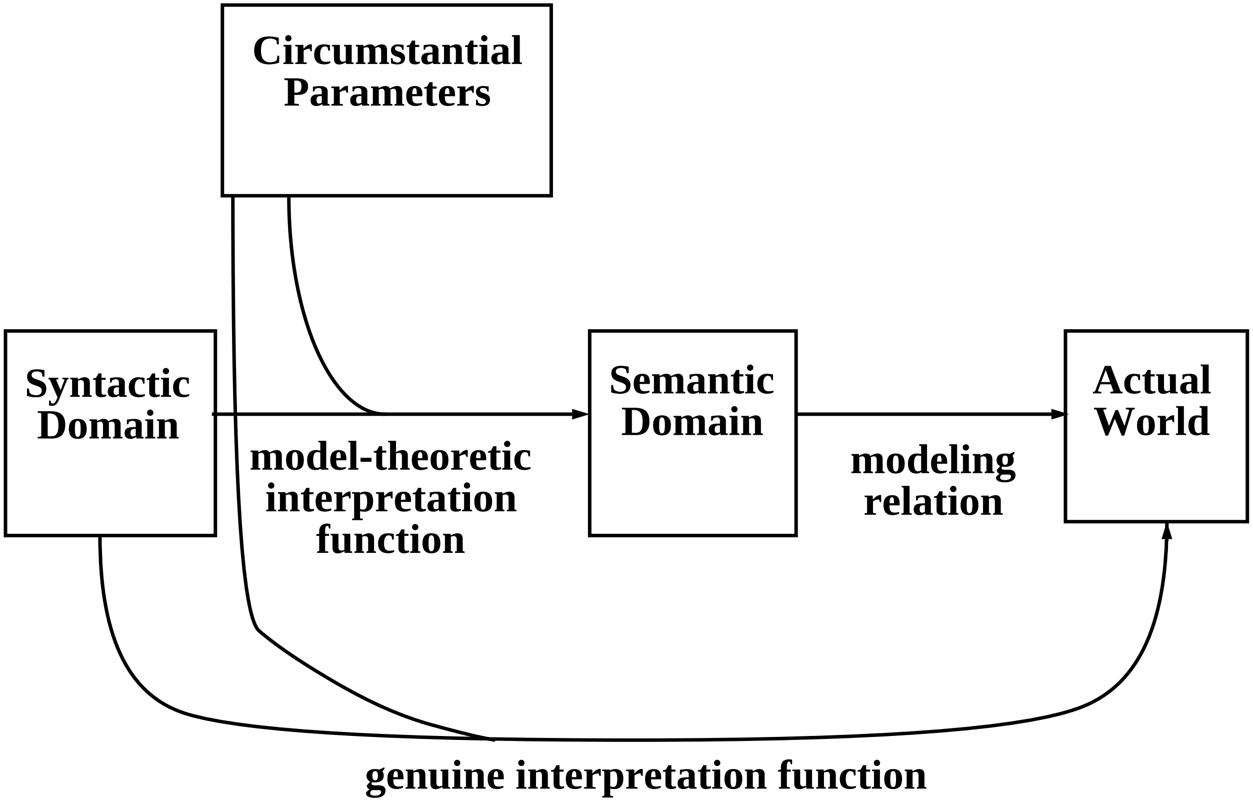

Circumstantial Parameters
Syntactic Domain
Semantic Domain
Actual World
model-theoretic interpretation function
modeling relation
genuine interpretation function

Figure 1.4: Smith's commutative semantic diagram.

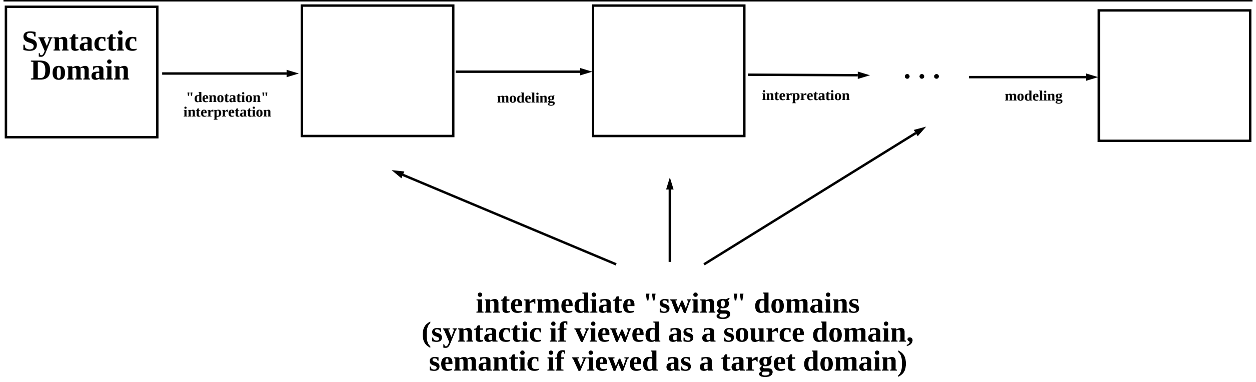

Syntactic Domain
"denotation" interpretation
modeling
interpretation
. . .
modeling
intermediate "swing" domains
(syntactic if viewed as a source domain,
semantic if viewed as a target domain)

Figure 1.5: Smith's correspondence continuum.

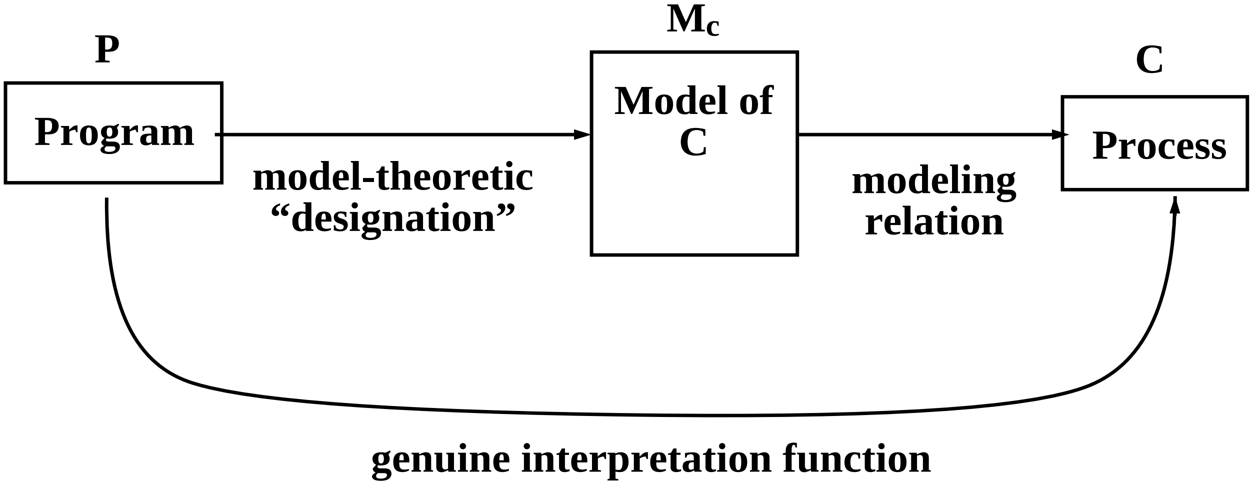


Figure 1.6: Smith’s view of programs and processes.

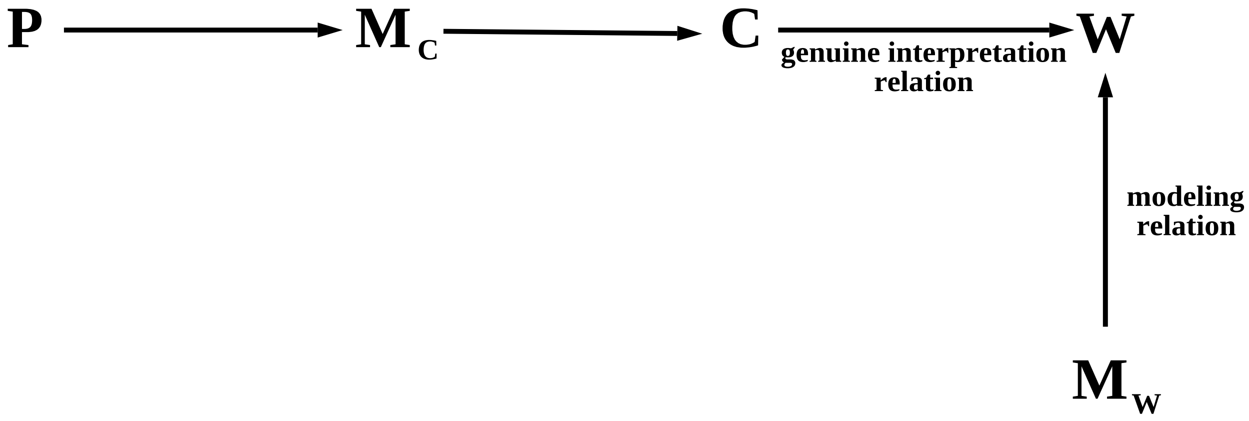


Figure 1.7: Smith’s view of programs and processes, elaborated.

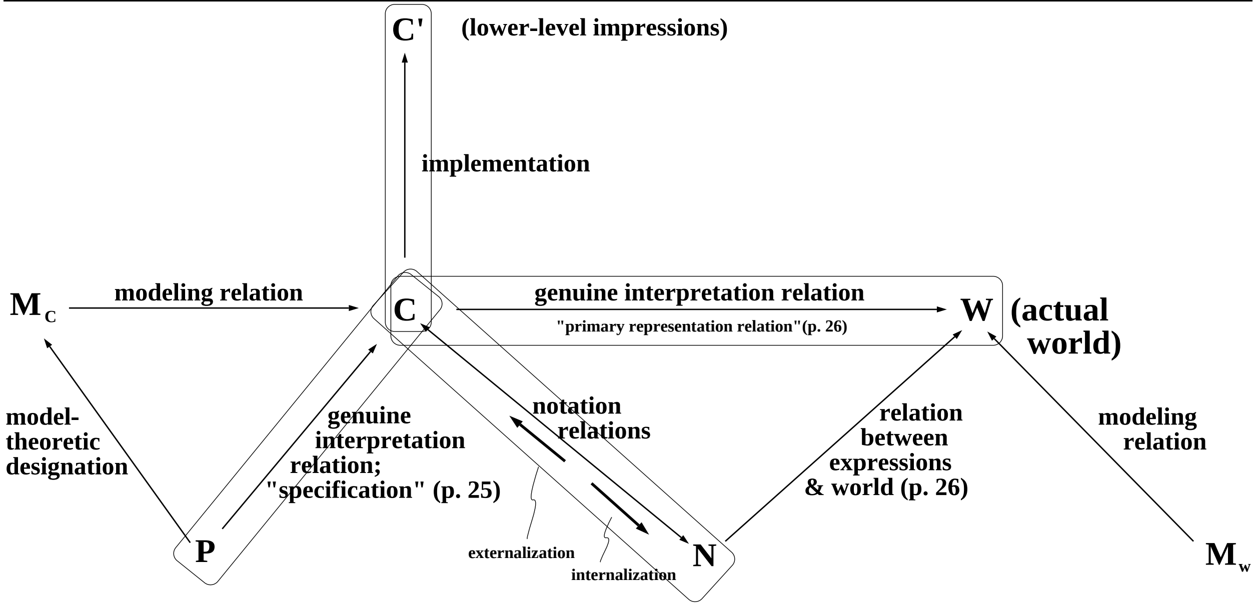

C'
(lower-level impressions)
implementation
M C
modeling relation
C
genuine interpretation relation
"primary representation relation"(p. 26)
W
(actual world)
model-theoretic designation
genuine interpretation relation; "specification" (p. 25)
notation relations
relation between expressions & world (p. 26)
modeling relation
P
externalization
internalization
N
M w

Figure 1.8: Smith's full picture.

Smith says that C's "impressions and behaviour" are "describe[d] ... in terms of the corresponding impressions or behaviour of" that machine (p. 22). But it is better to say that the description is in terms of the impressions and behavior of *a computational process* C′ of the lower-level machine. Yet these "two" processes are really the *same* (as I have argued in "Computer Processes and Virtual Persons" (Rapaport 1990)).

For another thing, there is a notation, N—a language for expressing C's impressions—with a pair of relations that "internalize" N into C and "externalize" C into N (p. 24). The notation N, as well as the process C, is also related to the actual world W, and, presumably, the diagram commutes. So the full picture is as shown in Figure 1.8. The implementation relation between C and C′, the notation relations between C and N, and the genuine interpretation relations between C and W and between C and P are the "genuine" ones—they are "causal" (p. 26). Process C is the semantic domain for "specification" (and P is its syntactic domain), and C is the syntactic domain for "primary representation" (and W is its semantic domain). Thus, C is what I've been calling a "swing" domain.

But Smith also takes C as the semantic domain for "notation" and "implementation". As for notation, surely N is the syntactic domain, so it's only "internalization", not "notation" in general—and certainly not "externalization"—for which C is the semantic domain (in the "classical" sense, of course; by my lights, what counts as syntactic or semantic depends on which is taken as antecedently understood). In the case of externalization, I would say that C is the syntactic, and N the semantic, domain: Expressions implement impressions in the physical medium of speech or writing. As for implementation, surely C is the syntactic domain and C′ is the semantic one; that is what implementation is all about (or so I shall argue in Chapter 1).

Some semantic relations, for Smith, are transitive; others aren't. The transitive ones are "modeling" relations; the others are "denotation" relations (p. 27). Consider, as he suggests, a photo (P_2) of a photo (P_1) of a ship (S). Smith observes that P_2 is not, on pain of use–mention confusion, a photo of S, but that this is "pedantic". Clearly, there are differences between P_1 and P_2: Properties of P_1 *per se* (say, a scratch on the negative) might appear in P_2 and be mistakenly attributed to S. But consider a photo of a map of the world (as in Figure 1.9, an ad for New York University that appeared in *The New York Times* (20 August

1991: D5)); the photo *could* be used as a map of the world. As Smith points out, the photo of the map isn't a map (just as P_2 isn't a photo of S). Yet *information* is preserved, so the photo can be *used as* a map (or: to the extent that information is preserved, it can be so used).

Another of Smith's examples is a document-image–understanding system, which has a knowledge representation of a digital image of a photo (cf. Srihari's system, example 3, above). What represents what? Does the knowledge representation represent the digitized image, or does it represent the photo? The practical value of such a system lies in the knowledge representation representing the photo, not the (intermediate) digitized image. But perhaps, to be pedantic about it, we should say that the knowledge representation does represent the digitized image even though *we* take it *as* representing the photo. After all, the digitized image is internal to the document-image–understanding system, which has no direct access to the photo. Of course, neither do we. Smith seems to agree:

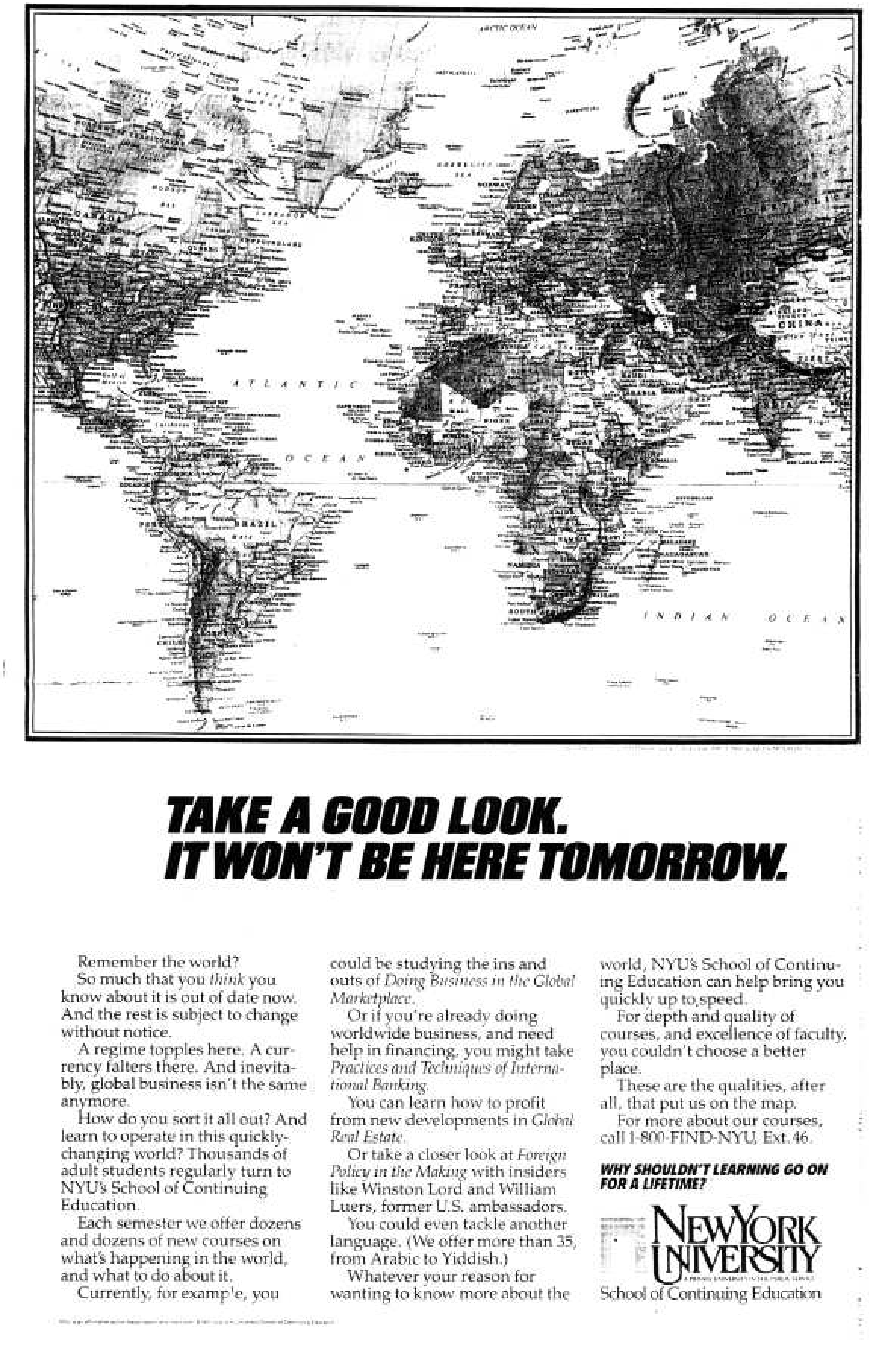
> The true situation ... is this: a given intentional structure—language, process, impression, model—is set in correspondence with one or more other structures, each of which is in turn set in correspondence with still others, at some point reaching (we hope) the states of affairs in the world that the original structures were genuinely about.
>
> It is this structure that I call the 'correspondence continuum'—a semantic soup in which to locate transitive and non-transitive linguistic relations, relations of modelling and encoding, implementation and realisation .... (Smith 1987: 29.)

But can one distinguish among this variety of relations? What makes modeling different from implementation, say? Perhaps one can distinguish between transitive and non-transitive semantic relations, but within those two categories, can useful distinctions really be drawn, say, between modeling, encoding, implementation, etc.? I think not. Perhaps one can say that there are "intended" distinctions, but (how) can these be pinned down? I think they can't. Perhaps one can say that it is the person doing the relating who decides, but is that any more than giving different names or offering external purposes? Indeed, Smith suggests (p. 29) that the only differences are individual ones.

He thinks, though, that not all "of these correspondence relations should be counted as genuinely semantic, intentional, representational" (p. 30), citing as an example the correspondences between (1a) an optic-nerve signal and (1b) a retinal intensity pattern, between (2a) the retinal intensity pattern and (2b) light-wave structures, between (3a) light-wave structures and (3c) "surface shape on which sunlight falls", and between (4a) that sunlit surface shape and (4b) a cat. He observes that "it is the cat that I see, not any of these intermediary structures" (p. 30). But so what? Some correspondence relations are not present to consciousness. Nonetheless, they can be treated as semantic.

Not so, says Smith: "correspondence is a far more general phenomenon than representation or interpretation" (p. 30). What is it to be "genuinely semantic" (p. 30)? Is it to be *about* something? But why *can't* we say that the retinal intensity pattern is "about" the light-wave structures? Or that the light-wave structures are "about" the sunlit surface shape? Isn't the relation between two of these purely physical processes one of information transfer (in either Shannon's (1949) or Dretske's (1981) sense)? If so, it is surely semantic. Note that it seems to be precisely when phenomena are information-theoretic that models of them *are* the phenomena themselves: Photos of maps *are* maps; models of minds *are* minds. (Perhaps it would be better to say that they are *instances of* the phenomena themselves (I owe this point to V. Kripasundar). Even better, they *can be* the phenomena (I owe this point to Kean Kaufmann)—this leaves open the question of when they *are*. As Kripasundar pointed out (personal communication), one gas can be modeled by another gas, yet this is not information-theoretic. Perhaps we should just say that for (non-trivial) information-theoretic phenomena, models are [or can be] the phenomena themselves.)

Smith proposes that for a correspondence to be semantic, it must be (1) "disconnected"—the representation and the represented must be disconnected—and (2) "registered"—representations represent the world in a certain way (p. 54n17). Disconnectivity seems related to the possibility of error (cf. p. 4: the level of sap in a maple tree is correlated with sugar production, but "sap can't be wrong"). But couldn't retinal intensity patterns be in error? And, anyway, why is error important? As for registration, surely my


ARCTIC OCEAN
ATLANTIC OCEAN
BRAZIL
CHINA
INDIAN OCEAN

**TAKE A GOOD LOOK.
IT WON'T BE HERE TOMORROW.**

Remember the world?

So much that you *think* you know about it is out of date now. And the rest is subject to change without notice.

A regime topples here. A currency falters there. And inevitably, global business isn't the same anymore.

How do you sort it all out? And learn to operate in this quickly-changing world? Thousands of adult students regularly turn to NYU's School of Continuing Education.

Each semester we offer dozens and dozens of new courses on what's happening in the world, and what to do about it.

Currently, for examp'e, you could be studying the ins and outs of *Doing Business in the Global Marketplace*.

Or if you're already doing worldwide business, and need help in financing, you might take *Practices and Techniques of International Banking*.

You can learn how to profit from new developments in *Global Real Estate*.

Or take a closer look at *Foreign Policy in the Making* with insiders like Winston Lord and William Luers, former U.S. ambassadors.

You could even tackle another language. (We offer more than 35, from Arabic to Yiddish.)

Whatever your reason for wanting to know more about the world, NYU's School of Continuing Education can help bring you quickly up to,speed.

For depth and quality of courses, and excellence of faculty, you couldn't choose a better place.

These are the qualities, after all, that put us on the map.

For more about our courses, call 1-800-FIND-NYU, Ext.46.

WHY SHOULDN'T LEARNING GO ON FOR A LIFETIME?

NEW YORK UNIVERSITY
School of Continuing Education

Figure 1.9: Ad for NYU, *New York Times* (20 August 1991: D5). Is this a map? A photo of a map? A reproduction of a photo of a map?

retinal intensity patterns only "import" part of the light-wave structures, which in turn "import" only part of the surface shape. This *aspectual* feature arising from partiality seems quite general and not limited to "genuine semantics".

Indeed, in his presentation of a general theory of correspondence between a domain and a co-domain, he says that "specific correspondence relations are defined between states of affairs in each domain—... between things *being a certain way* in one domain, and things *being a certain way* in the other" (p. 31). So the correspondences are between *aspects* of elements of the domain and co-domain; this seems to capture the "registration" feature. This interpretation of Smith's theory is supported by his noting that not all features of a domain element correspond to features of co-domain elements (p. 32). In fact, he says that it's necessary to pre-identify the states of affairs before specifying the correspondence relation, and he calls this process "registration" (p. 32).

I shall refrain from an analysis of his theory of correspondence (except to note that it bears comparison with the earlier and less-well-known theory of Apostel 1960)). What is important for my purposes is his claim that

> the correspondence continuum challenges the clear difference between "syntactic" and "semantic" analyses of representational formalisms . . . . . . [N]o simple "syntactic/semantic" distinction gets at a natural joint in the underlying subject matter. (Smith 1987: 38.)

Although he might be making the point that there can be no "pure" syntactic (or semantic) analyses—that each involves the other—his discussion suggests that the "challenge" is the existence of swing domains.

(The correspondence continuum plays a bit of havoc with (or: illuminates) the notion of compositionality:

> . . . when either or both domains are analysed mereologically—in terms of notions of part and whole—either or both ends of the correspondence can be defined *compositionally*, in the sense that what corresponds to (or is corresponded to by) a whole is systematically constituted out of what corresponds to (or, again, is corresponded to by) its parts. (Smith 1987: 33–34.)

That is, either or both ends of the correspondence can be "linguistic", as in Assumption 1. But note the oddity: It is the *domain* that is "compositional"; normally, one says that the (semantic) *relation* between the domains is compositional.)

Let's try an example. Let D and C be the domain and co-domain, let R be the correspondence relation between them, and suppose for now that R is a function. Suppose that D is mereologically analyzed. Let d_i be atomic elements of D, and let δ_j be operations that take sets of d_is and produce molecular elements Δ_k of D (that is, the d_is are "parts" of the Δ_ks, which are "wholes"). Next, suppose that $\mathrm{R}(\Delta_k) = c_i \in \mathrm{C}$, where $\Delta_k = \delta_j(d_1, \ldots, d_n)$. Normally, we would say that it is R that is compositional if $c_i = \mathrm{R}(\delta_j)(\mathrm{R}(d_1), \ldots, \mathrm{R}(d_n))$, that is, if R is computed by taking the $\mathrm{R}(d_i)$s (either base cases or computed recursively) and combining *them* by δ_j's image under R to produce c_i. So, for R to be compositional, in the ordinary sense, D must be mereological. Does this ordinary compositionality of R require C to be mereological? Not if the $\mathrm{R}(d_i)$ aren't "parts" of c_i (that is, of $\mathrm{R}(\Delta_k)$). Yet what Smith *says* is that one "end of the correspondence" (say, D) "can be defined compositionally, in the sense that what corresponds to . . . a whole [viz., c_i, which corresponds to Δ_k] is systematically constituted out of what corresponds to . . . its parts"; that is, c_i must be "systematically constituted out of" the $\mathrm{R}(d_i)$s. But that makes C mereological!

But perhaps I have it wrong; perhaps D is compositional in Smith's sense if "what . . . *is corresponded to by* a whole is separately constituted out of what . . . *is corresponded to by* its parts"—Δ_k is systematically constituted by the d_is. But this would still require c_i to be a *whole*, hence for C to be mereological. So, if D is mereological, so must C be, if Smith is to be taken literally.

I don't think he should. What he is suggesting, I think, is that there are two kinds of compositionality *of the correspondence relation* R: one in which R depends on D being mereological and one in which R depends on C being mereological. If both are mereological, then R could be compositional in two *prima facie* different senses. Examples, however, are not provided. I leave the details as an exercise for the reader.

1.7.4 The Gap, Revisited.

So we have a continuum, or at least a chained sequence, of domains that correspond to one another, each (except the last) understandable in terms of the next (or, occasionally, in terms of one further down the chain, with the intermediate domains being "invisible"). Yet where the last domain is the actual world, there is—as Smith has shown us—a gap between it and any model of it. Nonetheless, if that model of the world is in the mind of a cognitive agent—if it is *Cassie's* mental model of the world—then it was constructed (or it developed) by means of perception, communication, and other direct experience or direct contact with the actual world. In terms of Smith's three-link chain consisting of a part of the actual world (W), a set-theoretic model of it (M_W), and a linguistic description (in some program) of the model (D_{M_W}), what we have in Cassie's case is that her mental model of the world is simultaneously M_W and D_{M_W}. It is produced by causal links with the external world. Thus, the gap is, in fact, bridged (in this case, at least). Bridged, but not comprehended. In formalizing something that is essentially *in*formal, one can't *prove* (formally, of course) that the formalization is correct; one can only discuss it with other formalizers and come to some (perhaps tentative, perhaps conventional) agreement about it. Thus, Cassie can never check to see if her formal M_W really does match the informal, messy W. Thus, the gap remains. (And, once bridged, M_W is independent of W, except when Cassie interacts with W by conversing, asking a question, or acting. That is the lesson of methodological solipsism.)

It is time, now, to turn to Cassie's construction of M_W.

1.8 CASSIE'S MENTAL MODEL.

How does Cassie (or any (computational) cognitive agent, for that matter) construct her mental model of the world, and what does that model look like? I will focus on her language-understanding abilities—her mental model of a conversation or narrative. (For a discussion of how she might perceive visually, see the references to Srihari's system, cited in example 3, above.) Many of the details of Cassie's language-understanding abilities have been discussed in a series of earlier papers, with which familiarity is assumed.[9] Here, I will concentrate on two issues: a broad picture of how she processes linguistic input, and a consideration of the kind of world model she constructs as a result.

1.8.1 Fregean Semantics.

Frege wanted to divorce logic and semantics from psychology. In "On Sense and Reference" (1892), he tells us that terms and expressions (signs, or symbols) of a language "express" (*ausdrücken*) a "sense" (*Sinn*) and that to some—but not all—*senses* there "corresponds" (*entsprechen*) a "referent" (*Bedeutung*). Indirectly, then, expressions "designate" or "refer" (or fail to designate or refer) to a referent. Further, the sense is the "way" (*Art*) in which the referent is presented by the expression. Except for the mentalistic notion of an "associated idea", which he does not take very seriously, all of this is very objective or non-cognitive.

Without pretending to do Frege scholarship, I want to show how something exactly like this goes on in cognition, when Cassie—and, I submit, any natural-language–understanding cognitive agent—understands

[9] Shapiro 1982, 1989; Almeida & Shapiro 1983; Rapaport & Shapiro 1984, 1995; Bruder et al. 1986; Li 1986; Rapaport 1986a; Rapaport, Shapiro, & Wiebe 1986; Wiebe & Rapaport 1986, 1988; Almeida 1987, 1995; Peters & Shapiro 1987ab; Shapiro & Rapaport 1987, 1991, 1995; Peters, Shapiro, & Rapaport 1988; Rapaport 1988, 1991a; Rapaport, Segal, Shapiro, Zubin, Bruder, Duchan, Almeida et al. 1989; Rapaport, Segal, Shapiro, Zubin, Bruder, Duchan, & Mark 1989; Wyatt 1989, 1990, 1993; Peters & Rapaport 1990; Wiebe 1990, 1991, 1994; Yuhan 1991; Yuhan & Shapiro 1995; see also Neal 1981, 1985; Neal & Shapiro 1984, 1985, 1987; Neal, Thielman, et al. 1989.

language. It is really quite simple:

1. Cassie perceives (hears or reads) a sentence.

2. By various computational processes (namely, the augmented-transition-network parser with its attendant lexical and morphological modules, plus various modules for dealing with anaphora resolution, computing belief spaces and subjective contexts, etc.), she constructs a node (or finds an already constructed one) in the semantic network that is her mental model.

3. That node constitutes her understanding of the perceived sentence (cf. also Terry Winograd's SHRDLU (1972)).

Now, the procedures that take pieces of language as input and produce nodes as output are algorithms—*ways* in which the nodes are associated with the linguistic symbols. They are, thus, akin to senses, and the nodes are akin to referents (cf. Wilks 1972). Here, though, all symbols denote, even 'unicorn' and 'round square'. That is, if Cassie hears or reads about, say, a unicorn, she constructs a node representing her concept (her understanding) of that unicorn. Her nodes represent the things she has thought about, whether or not they exist—they are part of her "epistemological ontology" (Rapaport 1985/1986).

I hasten to point out that there is a very different correspondence one can set up between natural-language understanding and Frege's theory. According to this correspondence, it is the node in Cassie's mental model that is akin to a sense, and it is an object (if one exists) in the actual world to which that node corresponds that is akin to the referent. On this view, Cassie's unicorn-node represents (or perhaps is) the sense of what she read about; and, of course (unfortunately), there is no corresponding referent in the external world. Modulo the subjectivity or psychologism of this correspondence (Frege would not have identified a sense with an expression of a language of thought), this is surely closer in spirit to Frege's enterprise.

Nonetheless, I find the first correspondence illuminating. It shows how senses can be interpreted as algorithms that yield referents (a kind of "procedural semantics" (see, e.g., Winograd 1975, Smith 1982b)). It also avoids the problem of non-denoting expressions: If no "referent" is found, one is just constructed, in a Meinongian spirit (cf. Rapaport 1981).

The various links between thought and language are direct and causal. Consider natural-language generation, the inverse of natural-language understanding. Cassie has certain thoughts; these are private to her. (Except, of course, that I, as her programmer and a "computational neuroscientist" (so to speak), have direct access to her thoughts and can manipulate them "directly" in the sense of not having to manipulate them via language. That is, as her programmer, I can literally "read her mind" and "put thoughts into her head". But I ought, on methodological (if not moral!) grounds, to refrain from doing so (as much as possible). I *should* only "change her mind" via conversation.) By means of various natural-language–generation algorithms (including, perhaps, the inverse of (some of) her natural-language–understanding algorithms), she produces—directly and causally, from her private mental model—public language, utterances (or inscriptions). I hear (or read) these; this begins the process of natural-language *understanding*. By means of *my* natural-language–understanding algorithms, I interpret her utterances, producing—directly and causally—my private thoughts. Thus, I interpret another's private thoughts indirectly, by directly interpreting her public expressions of those thoughts, which public expressions are, in turn, her direct expressions of her private thoughts.[10]

The two direct links are both semantic interpretations. The public expression of Cassie's thoughts is a semantic interpretation (in our perhaps extended sense); it is, in fact, an "implementation" or physical "realization" of her thoughts. And my understanding of what she says is a semantic interpretation of her public utterances. Thus, the public communication language (Shapiro 1993) is a "swing domain".

[10] Cf. the quotation from Gracia, §??.

1.8.2 The Nature of a Mental Model.

> Metaphysically the basic fact is that we have NO access to an external point of view. All reference is from *our*, *one's* point of view. (As is well known, here lies the kernel of Kant's Copernical Revolution.) (Castañeda 1989d: 35.)

Cassie's mental model of the world (including that part of the world consisting of utterances expressed in the public communication language) is expressed in her language of thought. That is, the world is modeled, or represented, by expressions of her language of thought. Her mental model consists, if you will, of sentences of that language of thought (which, for the sake of concreteness, I am taking to be SNePS). There may, of course, be more: for instance, mental imagery (corresponding to all sensory modalities—thus, mental visual images, mental auditory images, etc.). But since Cassie can think and talk about these images, they must be linked to the part of her mental model constructed via natural-language understanding (as suggested in Srihari 1991ab). Hence, we may consider them part of an extended language of thought that allows such imagery among its terms (and, perhaps, propositions). This extended language of thought, then, is propositional with direct connections to imagistic representations. However, Philip Johnson-Laird (1983) suggests that mental models have a somewhat different structure. Let us consider the nature of mental models in the context of Jon Barwise and John Etchemendy's (1989) discussion of the role of model-theoretic semantics in cognitive science.

In the study of thought and language, as contrasted with "most of what science sets out to explain ... there seems to be an entirely new type of property—'aboutness' or 'semantic content'—in need of explanation. This property is sometimes called the 'intentionality' of language and thought" (Barwise & Etchemendy 1989: 207). The notion of "content" is both vague and ambiguous. It is vague insofar as there is no clear, well-established definition of it, but this is true even for so well-entrenched and familiar a term as 'belief'. More serious is its ambiguity. Etymologically, it ought to be something "contained" within a piece of language or thought, and historically that was sometimes the case. Witness, say, Twardowsky's use of the term to mean something that is "completely within the [thinking] subject" (Twardowski 1894: 1–2) and that even "objectless" ideas (that is, ideas of non-existents) have (Twardowski 1894: 18). Often, though, it is used to mean something external to thought and language—indeed, something located in the external world, to which thought or language refers. Considering it as a synonym for 'intentionality', of course, does not disambiguate it, though it does favor the external interpretation, since intentionality as introduced by Brentano (1874) is the "directedness" of a mental act to an (external) object, to be contrasted with the content of the act.

In Chapter **??**, I posed as the central concern of this essay how we have knowledge of the semantics of our language. Barwise and Etchemendy take this as "a task for the cognitive scientist" (p. 209). It is the challenge posed by John Searle's Chinese Room Argument: How could Searle-in-the-room come to know the semantics of the Chinese squiggles? What is the Chinese-Room Argument and who is Searle-in-the-room? Searle has offered a thought experiment that has come to be called the Chinese-Room Argument (Searle 1980).

> In this experiment, Searle, who knows neither written nor spoken Chinese, is imagined to be locked in a room and supplied with an elaborate algorithm written in English that tells him [*de re*] how to write Chinese characters in response to other Chinese characters. Native Chinese speakers are stationed outside the room and pass pieces of paper with questions written in Chinese characters into the room. Searle uses these symbols, otherwise meaningless to him, as input and—following only the algorithm—produces, as output, answers written in Chinese characters. He passes these back outside to the native speakers, who find his "answers ... absolutely indistinguishable from those of native Chinese speakers" [(Searle 1980: 418)]. The argument that this experiment is supposed to support has been expressed by Searle ... as follows:
>
> > ... I still don't understand a word of Chinese and neither does any other digital computer because all the computer has is what I have: a formal program that attaches no meaning, interpretation, or content to any of the symbols.

> [Therefore,] . . . no formal program by itself is sufficient for understanding . . . [(Searle 1982: 5.)

(Rapaport 1986b: 7–8.)[11]

So, the question is: How could Searle-in-the-room know what the symbols he manipulates are about? One question that has been left open in the debate is whether Searle-in-the-room even knows what their *syntax* is. Could he come to know the syntax (the grammar)? Not, presumably, just by having, as Searle suggests, a SAM-like program (that is, a program for global understanding of a narrative; cf., e.g., Schank & Riesbeck 1981); a syntax-learning program is also needed (cf. §**??**, above). But we can assume that Searle-in-the-room's instruction book includes this (there has been, after all, lots of work on this topic; cf. Hedrick 1976; Wolff 1978, 1982; Berwick 1979, 1980; Langley 1980, 1982).

Given an understanding of the syntax, how can semantics be learned? In two ways, at least: ostensively and lexically. The meaning of some terms is best learned ostensively, or perceptually: We must see (or hear, or otherwise experience) that which the term refers to. This ranges from terms for such archetypally medium-sized physical objects as 'cat' and 'cow', through 'red' (cf. Jackson 1986) and 'internal combustion engine', to such abstractions as 'democracy' and 'love' (cf. how Helen Keller learned 'love' and 'think'; see §**??**).

But the meaning of many, perhaps most, terms is learned "lexically", or linguistically. Such is dictionary learning. But equally there is the learning, on the fly, of the meaning of new words from the

[11] In Searle's own words:

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore . . . that I know no Chinese . . . To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. Now suppose that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch "a script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions," and the set of rules in English that they give me, they call "the program." . . . [I]magine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. . . . Let us also suppose that my answers to the English questions are . . . indistinguishable from those of other native English speakers . . . . From the external point of view—from the point of view of someone reading my "answers"—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.
>
> Now the claims made by strong AI are that the programmed computer understands the stories and that the program in some sense explains human understanding. . . .
>
> . . . [I]t seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. For the same reasons, . . . [a] computer understands nothing of any stories . . . .
>
> . . . [W]e can see that the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding. But does it even provide a necessary condition . . . ? One of the claims made by the supporters of strong AI is that when I understand a story in English, what I am doing is exactly the same . . . as what I was doing in manipulating the Chinese symbols. . . . I have not demonstrated that this claim is false . . . . As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. . . . [W]hatever purely formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything. (Searle 1980: 417–418.)

linguistic contexts in which they appear. This can be thought of algebraically: "the appearance of a word in a restricted number of settings suffices to determine its position in the language as a whole" (Higginbotham 1985: 2): If 'vase' is unknown, but one learns that Tommy broke a vase, then one can compute that a vase is that which Tommy broke (Ehrlich 1995). Initially, this may appear less than informative, though further inferences can be drawn: Vases, whatever they are, are breakable by humans, and all that that entails. As more occurrences of the word are encountered, the "simultaneous equations" (Higginbotham 1989: 469) of the differing contexts, together with background knowledge and some guesswork, help constrain the meaning further, allowing us to revise our theory of the word's meaning. Sooner or later, a provisionally steady state is achieved (pending future occurrences). (For more details, see §**??**. Cf. Rapaport 1981; Ehrlich & Rapaport 1992, 1993, 1995; Ehrlich 1995.)

Both methods are contextual. For ostension, the context is physical and external—the real world (or, at least, our perception of it); this is the "wide context" of Rapaport 1981. For the lexical, the context is linguistic (the "narrow context" of Rapaport 1981). Ultimately, the context is mental and internal: The meaning of a term represented by a node in a semantic network is dependent on its location in—that is, the surrounding context of—the rest of the network. (Cf. Quine 1951; Quillian 1967, 1968; Quine & Ullian 1978; Hill 1994, 1995.) Such holism has a long and distinguished history. It also has had its share of distinguished but obscure incarnations (for example, the Hegelian Absolute, parodied so nicely in F. C. S. Schiller's *Mind!* (1901)) and its share of skeptics (most recently, Fodor and Lepore (1992)). It certainly appears susceptible to charges of circularity (cf., for example, Harnad 1990), though perhaps a chronological theory of how the network is constructed can help to obviate that: Granted that the meaning of 'vase' (for me) may depend on the meaning of 'breakable' *and vice versa*, nonetheless I learned the meaning of the latter first; so it can be used to ground the meaning of the former (for me). Holism, though, has benefits: The meanings of terms get enriched, over time, the more they—or their closest-linked terms in the network—are encountered.

This ramifies upwards. In the preliminary note-taking research for this book, certain themes constantly reappeared in various contexts, each appearance enriching the others. In writing, however, one must begin somewhere—writing is a more or less sequential, not a parallel or even holistic, task. (I suppose hypermedia might implement a holistically written text.) Though this is the first appearance of holism in the essay, it was not the first in my research, nor will it be the only one (see Ch. **??**).

Understanding, we see again, is recursive. Each time we understand something, we understand it in terms of all that has come before. Each of those things, earlier understood, were understood in terms of what preceded them. The base case is, retroactively, understandable in terms of all that has come later.

> 1) The classics are the books of which we usually hear people say: "I am rereading . . . " and never "I am reading . . . ." [. . .]
>
> There should therefore be a time in adult life devoted to revisiting the most important books of our youth. Even if the books have remained the same (though they do change, in the light of an altered historical perspective), we have most certainly changed, and our encounter will be an entirely new thing.
>
> Hence, whether we use the verb "read" or the verb "reread" is of little importance. Indeed, we may say:
>
> 4) Every rereading of a classic is as much a voyage of discovery as the first reading.
>
> 5) Every reading of a classic is in fact a rereading. (Calvino 1986: 19.)

But initially, the base case was understandable solely in terms of itself (or in terms of "innate ideas" or some other mechanism—we will return to this later; cf., also, Hill 1994, 1995 on the semantics of base nodes in SNePS).

But *is* "knowledge of the semantics" achieved by speakers? If this means knowledge of the relations between word and thing, and if it means that in such a way that such knowledge requires knowledge of *both* the words (syntactic knowledge) *and* the things, then: No. For we can't have (direct) knowledge of the things. This is Smith's gap. It also means, by the way, that ostensive learning is really mental and internal, too: I learn what 'cat' means by seeing one, but really what's happening is that I have a mental representation of that which is before my eyes, and what constitutes the ostensive meaning is a (semantic) link that is established between my internal node associated with 'cat' and the *internal* node that represents what is before my eyes.

Thus, "knowledge of the semantics" means (1) knowledge of the relations *between* those of our concepts that are linguistic and those of our concepts that are "purely conceptual", that is, that correspond to, or are caused by, external input, and (2) knowledge of the relations *among* our purely linguistic concepts. The former (1) is "semantic", the latter (2) "syntactic", as classically construed (Morris 1938). Yet, since the former concerns relations among our internal concepts (cf. Srihari 1991ab), it, too, is syntactic. (The first time you heard me say this, you either found it incomprehensible or insane. By now, it should be less of the former, if not of the latter, since its role in the web of my theory should be becoming clearer.)

Barwise and Etchemendy conflate such an internal semantic theory with a kind of external one, identifying "*content of a speaker's knowledge* of the truth conditions of the sentences of his or her language" with "*the relationship between sentences and non-linguistic facts* about the world that would support the truth of a claim made with the sentence" (p. 220, my italics). I take "the content of a speaker's knowledge of ... truth conditions" to involve knowing the relations between linguistic and non-linguistic *internal* concepts. This is the internal, Cassie-approach to semantics. In contrast, giving an "account of the relationship between sentences and non-linguistic facts" (p. 220) is an *external* endeavor, one that *I* can give concerning Cassie, but not one that *she* can give about herself. This is because *I* can take a "God's-eye", "third-person" point of view and see both Cassie's mind and the world external to it, thus being able to relate them, whereas she can only take the "first-person" point of view.

There are, however, some limitations on the third-person point of view:

1. A "third person" can only have direct access to a cognitive agent's mind in the case of a *computational* cognitive agent much as Cassie, not in the case of an ordinary human being. (At least, such is the state of affairs now; perhaps in the forthcoming golden age of neuroscience, my (current) access to Cassie's mind—my ability to literally look at her mind and literally change it in a direct fashion (not indirectly via language, perception, or inference)—will not differ significantly from such a golden-age neuroscientist's access to mind.)

2. More importantly, a "third person" cannot, in fact, have direct access to the external world. So what the third person is *really* comparing (or finding correspondences between) is Cassie's concepts (better: the third person's *representations* of Cassie's concepts) and the third person's *own concepts* representing the external world. That is, the third person *can* establish a semantic correspondence (in the classic sense) between two domains. From the third person's point of view, the two domains are the syntactic domain consisting of Cassie's concepts and the semantic domain of the external world. But in fact, the two domains are *the third person's representations* of Cassie's concepts and *the third person's representations* of the external world. These are both *internal* to the third person's mind! And internal relations, even though structurally *semantic*—that is, even though they are correspondences between two domains—are fundamentally *syntactic* in the classic sense: They are relations *among* (two classes of) symbols in the third person's language of thought.

What holds for the third person holds also for Cassie. Since she doesn't have direct access to the external world either, she can't have knowledge of "real" semantic correspondences. The best she can do is to have a correspondence between certain of her concepts and her representations of the external world. What might her "knowledge of truth conditions" look like? As a first suggestion, when she learns that Lucy is rich, she builds the network shown in Figure 1.10. (Linearly abbreviated: M2 = B1 is named 'Lucy'; M4 = B1 is rich). Thus, Cassie might think to herself something like: "My thought that $\mathtt{LUCY}\overset{\mathtt{M2!}}{\wedge}\mathtt{B1}\overset{\mathtt{M4!}}{\wedge}\mathtt{RICH}$ is true iff ($\exists x \in$ external world)[x = Lucy & x is rich]". This is purely syntactic, since both sides of the biconditional are expressed in Cassie's language of thought. (It would require, for its full development, (1) an internal truth predicate (cf. Maida & Shapiro 1982, Neal 1985, Neal & Shapiro 1987), (2) an existence predicate (cf. Hirst 1989, 1991), (3) a duplication of the network (but perhaps not: by the Uniqueness Principle

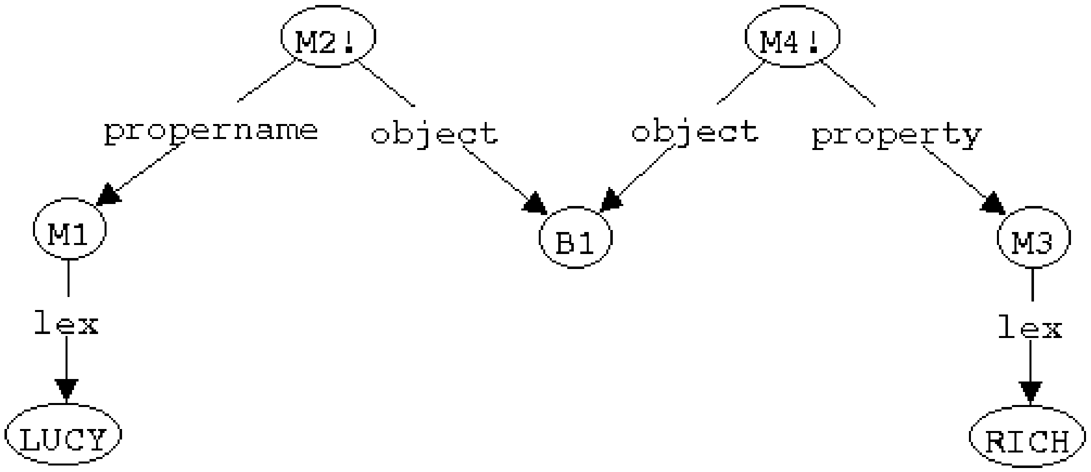


Figure 1.10: Cassie's belief that Lucy is rich.

(Maida & Shapiro 1982, Shapiro & Rapaport 1987), this network should—and could—be re-used), and (4) a biconditional rule asserting the equivalence (see Figure 1.11 for a possible version). Thus, the best Cassie can do is to have a coherence theory of truth: coherence among her own concepts.

Barwise and Etchemendy observe that "[t]o provide a rigorous analysis of this dependence [of the truth of a sentence on typically non-linguistic states of affairs], model-theoretic semantics first develops some machinery for *representing* these non-linguistic states of affairs" (p. 220, my italics). Granted, the *truth* value of a sentence depends on non-linguistic, *external* states of affairs. But note the move that Smith (1985) has sensitized us to: using a *representation* of these external states of affairs, which itself demands a semantic theory—a correspondence between the model and the world. *We* represent external objects by internal nodes, so they play the same role that set-theoretical models do. So model-theoretic semantic *techniques* are the same as (or are applicable to) the relation between what might be called "linguistic" nodes (for example, M4 in the example of Figure 1.11) and "non-linguistic" nodes (for example, P3). So that relationship is *both* semantic (model-theoretically) *and* syntactic (since it consists of relations among symbols).

The model muddle is not far away: "we introduce the notion of model w of the world. Because our [toy] language is designed for use in talking about the solar system, we could think of these models as mathematical models of the solar system, much as an orrery is a physical model of the solar system" (p. 220). Barwise and Etchemendy's use of 'model' is such that a set-theoretic structure is a model *of the world*, in the sense of a mathematical model. Normally, I think of model-theoretic models such as w as models *of the language*. Clearly, w is a swing domain: Let us say that it is both a model *of* the world and a model *for* the language.

What, by the way, is w's ontological status? Is it a "thing" consisting of mathematical structures, or is it a *linguistic* entity? I have always taken mathematical models to be linguistic, but perhaps this is merely my formalistic tendencies showing their face—mathematics seen as a *language* (syntax) rather than as that which the language is *about* (semantics). Of course, there *is* a language in any case, so if w is set-theoretic in the semantic sense (a "thing", rather than a linguistic entity), we *still* neeed a language to talk about the sets. So there are, then, *two* languages (or syntactic domains in the classical sense):

1. the "toy" language to talk about the solar system, in Barwise and Etchemendy's case, or language

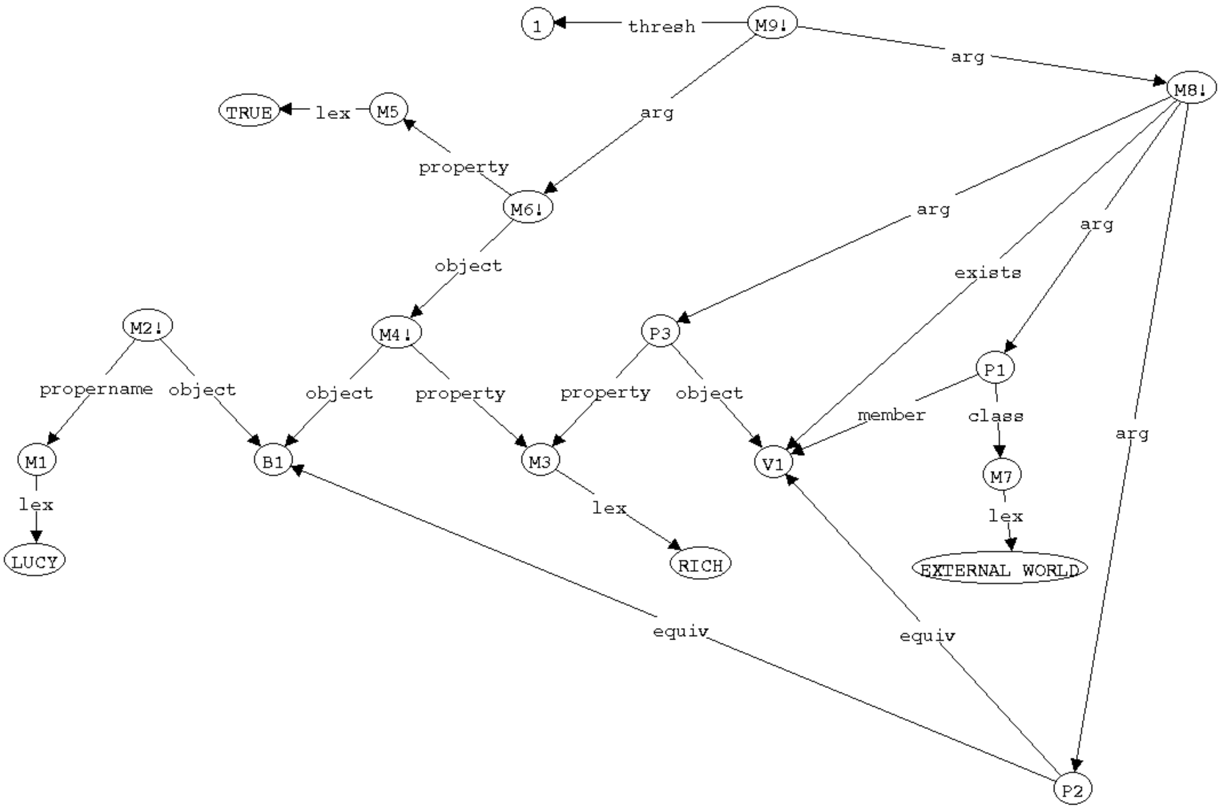


Figure 1.11: A biconditional rule (M9) asserting the equivalence of M6 = that Lucy is rich is true, and M8 = something in the external world is Lucy and is rich. More fully:

M6 = M4 is true;
M8 = $\exists$V1[P1 & P2 & P3];
P1 = V1 $\in$ external world;
P2 = V1 $\equiv$ B1;
P3 = V1 is rich;
M9 = M6 iff M8 (more precisely, if at least one of M6, M8 is true,then both are;
see Shapiro 1979, Shapiro & Rapaport 1987 for the semantics of thresh).

(Note that I've omitted the truth condition for M2.)

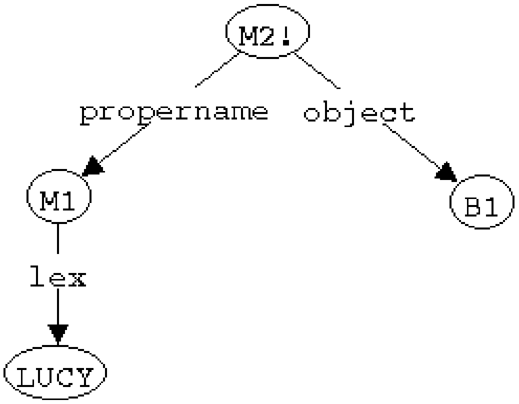


Figure 1.12: Cassie's belief that someone (B1) is named 'Lucy'.

simpliciter in the general case, and

2. the mathematical language to talk about w

and *two* ontological structures (or semantic domains in the classical sense):

3. the mathematical model w and
4. the real world (in Barwise and Etchemendy's case, the solar system).

On my view, two of the domains in this correspondence continuum, namely (2) and (3), are swing domains. Of course, this is all from a third-person point of view. From Cassie's first-person point of view, there are merely two languages: the internal, lexical, linguistic nodes and internal, non-linguistic nodes. The world, both real and mathematical, is inaccessible to her directly.

Digression. Now—a word to the reader. What follows is (a) pure open-ended speculation at this state and (b) probably only of interest to SNePS hackers. So, unless you fall into that category, you can ignore what follows. I'll let you know when you should start paying attention again. (See **Return from digression**, below.) Let's revise our first attempt at providing truth conditions for Cassie. Barwise and Etchemendy offer various semantic clauses (pp. 223ff) that we can mimic for Cassie. For instance, where t is a name, Barwise and Etchemendy say that the "interpretation" of t—$f(t)$—is its "denotation" in w under an assignment, g, of values to variables—$den(t, w, g)$. For Cassie, we can ignore g.

Suppose Cassie believes that someone is named 'Lucy' (see Figure 1.12). Recall that t is a name and that w is a model of the world (hence, w is *internal* to Cassie's mind). Presumably, then, the Barwise and Etchemendy domain of w, D^w, will be the set of non-linguistic nodes. Now, what is t? Is it M1, or is it the LUCY-node? If the latter, then perhaps $den(t, w) = f(t) = \mathtt{M1}$. If so, what's the relation between M1 and B1? If $t = \mathtt{M1}$, on the other hand, then $f(t) = \mathtt{B1}$; but then what's the relation between LUCY and M1?

Let's try a different approach. If we're really concerned with the semantics of *language*, then we need to consider Cassie's internal representations *of language*—internal representations of *sentences*, not beliefs. The internal representations of sentences, then, can correspond (both semantically, in the classical sense, and syntactically, since all is symbol manipulation) to her beliefs. We can use the representations of

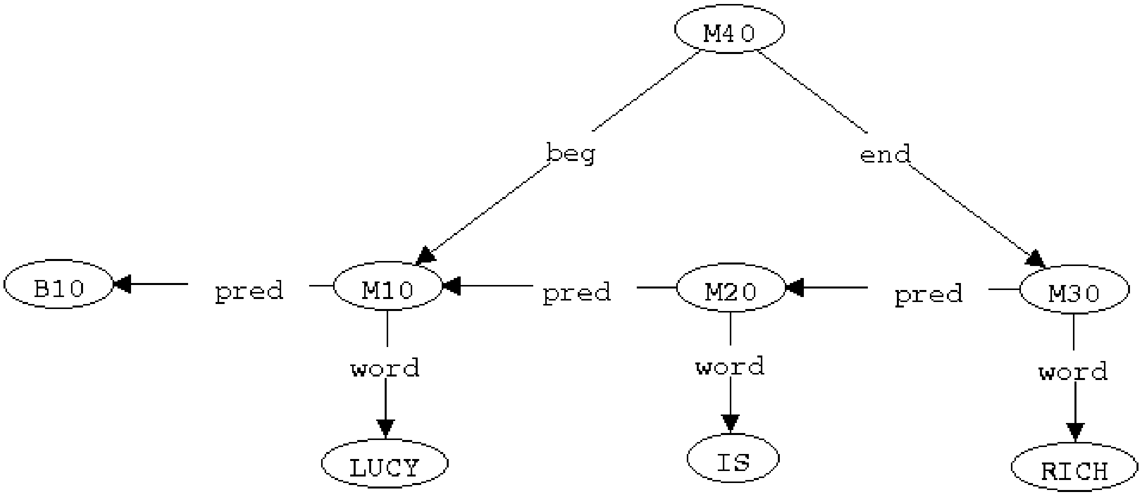


Figure 1.13: Cassie's representation of the *sentence* 'Lucy is rich' (roughly, `M40` = the sequence of words beginning with 'Lucy' and ending with 'rich'; after Neal & Shapiro 1987: 63).

Jeannette Neal (1981, 1985; Neal & Shapiro 1984, 1985, 1987). (On this view, t = `LUCY` (not `M1`).) Let's take a simple sentence: 'Lucy is rich'. Let Cassie's internal representation of this *sentence*—*qua* sentence—be as in Figure 1.13. Now, her *understanding*—her semantic interpretation—of that sentence is the belief shown in Figure 1.10. Then:

$$f(\texttt{M40}) = \texttt{M2}\ \&\ \texttt{M4}\ \text{(or, perhaps, just M4?)}$$

$$f(\texttt{LUCY}) = \texttt{M1}$$

$$f(\texttt{RICH}) = \texttt{M3}, \text{etc.}$$

And/or perhaps:

$$f(\texttt{M10}) = \texttt{B1}$$

$$f(\texttt{M30}) = \texttt{M3}, \text{etc.}$$

Question: Is the `LUCY` node dominated by `M1` the same node as the `LUCY` node dominated by `M10` (it should be, by the Uniqueness Principle), or is it the same node as `M10` itself?

Return from digression. OK; calling all non–SNePS-hackers. I'm finished exploring the nitty-gritty details. The important point is not the details I speculate on above, but that there *is* a way to have this kind of *internal* semantics (cf., also, Srihari 1991b; Lammens 1994, Ch. 3 and §7.4).

So, the picture (Fig. 1.11) we have of Cassie's mental model of the world (including utterances) is, in part, this: If Cassie hears or reads a sentence, she constructs a mental propositional representation of the sentence, *and* she constructs a mental representation of the state of affairs expressed by that sentence. These will be linked by a Tarski-like truth-biconditional asserting that the belief (`M4`) is true (`M6`) iff the representation of the state of affairs (`M8`) is believed (`M8!`). If Cassie sees something, she constructs a mental representation of it (in, say, Srihari-like notation), *and* she constructs a mental propositional representation of the state of affairs she sees. These will be linked in ways extrapolatable from Srihari (1991b, etc.). These networks, of course, are not isolated, but embedded in the entire network that has been constructed so far.

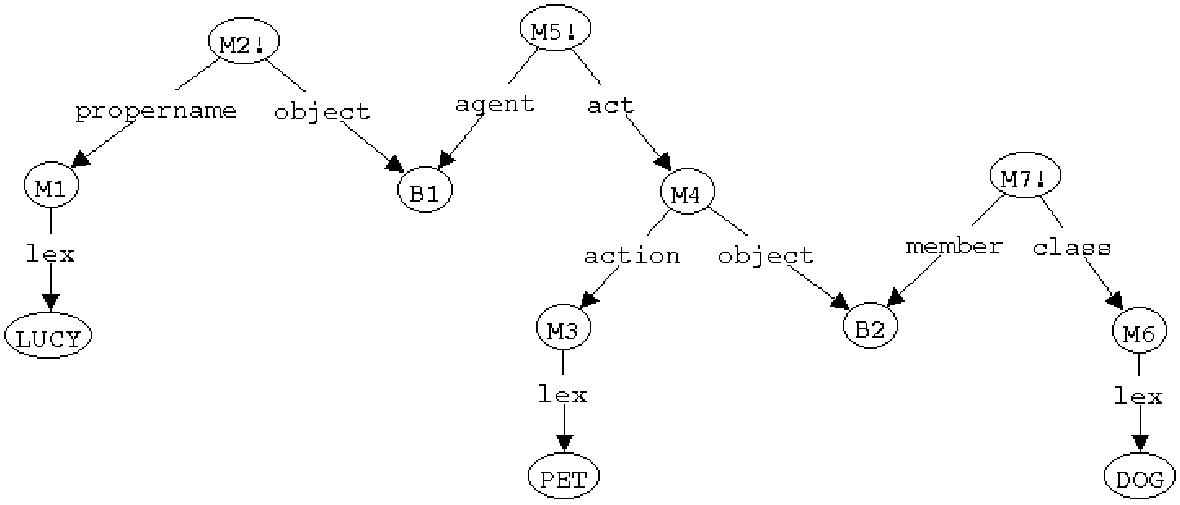


Figure 1.14: Cassie's belief that Lucy pets a dog:
M2 = B1 is named 'Lucy';
M7 = B2 is a dog;
M5 = B1 pets B2.

What is newly perceived is understood in terms of all that has gone before. This is purely syntactic, since both sides of the biconditional are expressed in Cassie's language of thought. Thus, the best Cassie can do is to have a theory of truth as coherence among her own concepts.

We now have enough background to return to Johnson-Laird's mental models, which differ in the details of representational notation (that is, in the language of thought) as well as in inference mechanisms. Since the latter, however, are dependent on the former, let us concentrate on the differing languages of thought. According to Barwise and Etchemendy, Johnson-Laird's

> [m]ental models are taken to be similar to mathematical models in two respects. First, as with our mathematical models they are taken to represent the world in a fairly direct "structural" way. This is why they are called mental "models" rather than, say, mental "sentences". (p. 227.)

Now, for Cassie, the appropriate comparison is to be made with what I've been calling 'non-linguistic nodes', that is, most of a typical SNePS network except for Neal-like linguistic structures and lex nodes. (By 'lex nodes', I mean the nodes at the heads of lex arcs. lex arcs emanate from nodes representing concepts expressed in the English lexicon by the word at the head of the lex arc. See Shapiro 1982, Rapaport 1988 for details. In Rapaport 1988, I suggested the use of pic arcs that would link concepts with visual images; a version of these were implemented in Srihari 1991b.) Nonetheless, such "non-linguistic" networks *are* sentences of a mental language—as are Johnson-Laird's representations: They have a formal syntax. Do SNePS nodes represent in a "direct, structural way"? Not quite: They *are* more language-like than Johnson-Laird's representations.

On the other hand, to represent the *state of affairs* of, say Lucy petting a dog is to represent that state of affairs as having the *structure* shown in Figure 1.14. It consists of an agent and an act; the agent is represented by a structureless base node (B1) (but we can assert things about it, for example, that it is named 'Lucy' (M2), that it is a person (not shown), etc.). The act has the following (sub)structure: It consists of an

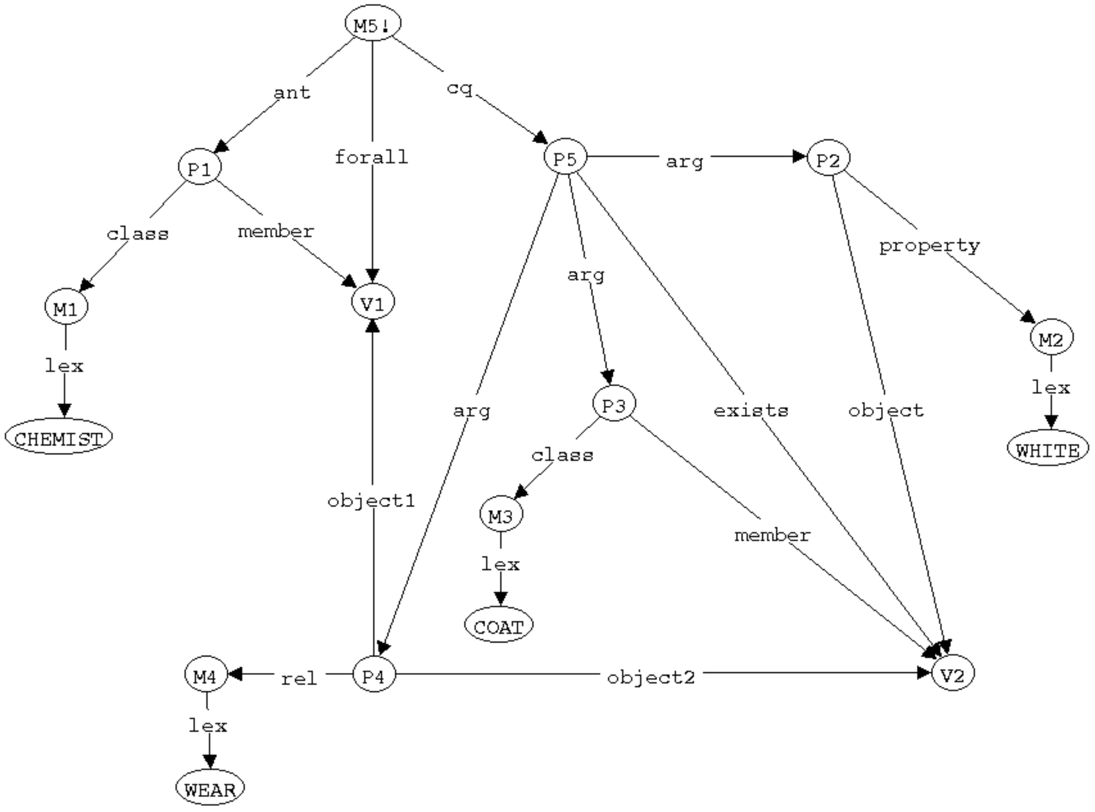


Figure 1.15: Cassie's belief that all chemists wear white coats:
M5 = ∀V1[P1 → P5];
P1 = V1 is a chemist;
P5 = ∃V2[P2 & P3 & P4];
P2 = V2 is white;
P3 = V2 is a coat;
P4 = V1 wears V2.

action (M3) and an object (B2); each of these is structureless, though each can have things asserted about it, for example, that the petted thing is a member of the class *dog* (M6), that petting is a physical activity (not shown), etc. (The action is structured only in the sense of being expressed by a particular lexical item.) So, our nodes *are* models in a Johnson-Laird–like sense, though the proposed structure is different.

The difference really shows up in quantified (especially numerically quantified) sentences such as 'All chemists wear white coats'. For Johnson-Laird, the structure is: lots of chemist-models, all of which are models of white-coat–wearers. For SNePS (see Figure 1.15), the structure is: a rule node (M5) consisting of a universally quantified arbitrary item (V1), an antecedent state of affairs P1 (actually, a *pattern* for a state of affairs), and a consequent state of affairs (pattern P5); the antecedent says that the arbitrary item is a member of a class M1 (expressed in English by 'chemist'); and the consequent consists of a rule node P5 (actually, a pattern for a rule) consisting of an existentially quantified arbitrary item (V2) and a conjunction of three patterns: P4, which represents that the first arbitrary item bears the relation M2 (expressed by 'wear') to the other arbitrary item, V2, i.e., that which is worn by the first arbitrary item; P2, which represents that V2 has property M3 (expressed by 'white'); and P3, which represents that V2 is a member of the class M4,

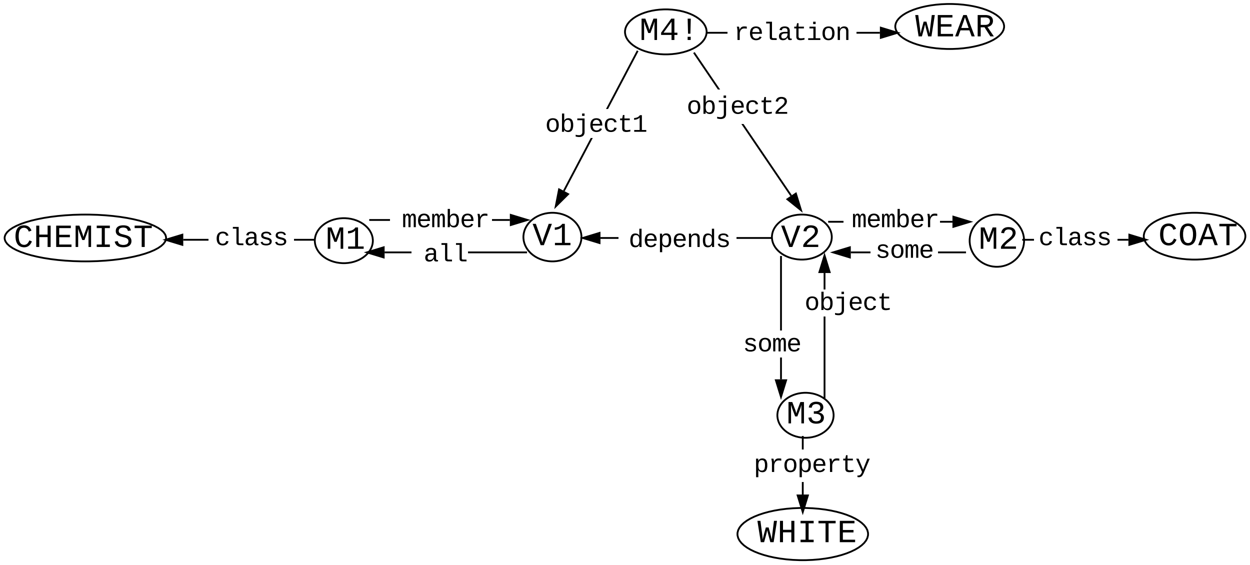


Figure 1.16: ANALOG representation of 'All chemists wear white coats':
M4 = V1 wears V2.
M1 = V1 is an arbitrary member of the class of chemists.
M2 = V2 is a member, depending on V1, of the class of coats and which is white.

expressed by 'coat'.

Syed Ali's ANALOG system (Ali 1994, 1995; Ali & Shapiro 1993; see Figure 1.16) uses a different SNePS representation that consists of an arbitrary chemist (V1) that wears a white coat (V2). This is a "prototype" approach rather than a Johnson-Laird–like "exemplar" approach. Note that both sorts of SNePS representations are more like the "situations" of Situation Semantics (cf., for example, Barwise & Perry 1983) or the discourse representation structures of Discourse Representation Theory than they are like a Johnson-Laird mental model. But all of them are mental models of the world. They are, thus, *semantic* interpretations of the world and of language, as rich and robust as you please. But they are expressed in formal languages of thought, so they are *syntactic* symbol systems.

1.8.3 The Psychological (and Biological) Reality of Mental Models.

All of this is fine as far as it goes, and much the same sort of thing can be said for representational connectionist systems—they, too, are mental models of the world, though the language of thought is radically different in syntax. But these are all computational models. Do *we* work that way? Antonio Damasio has provided some evidence that we do:

> Human experiences as they occur ephemerally in *perception* ... are based on the **cerebral representation** of concrete external entities, internal entities, abstract entities, and events.
>
> Such representations are **interrelated by combinatorial arrangements** so that their internal action in recall and the order with which they are attended, permits them to unfold in a "sentential" manner. **Such "sentences" embody semantic and syntactic principles.** (Damasio 1989a: 44; his italics, my boldface.)

If this isn't a language of thought, what is? There's more:

> Because feature-based fragments are recorded and reactivated in sensory and motor cortices, the **reconstitution** of an entity or event so that it resembles the original experience depends on the **recording of the combinatorial arrangement** that conjoined the fragments in perceptual or recalled experience. The record of each unique combinatorial arrangement is [what Damasio calls] the binding code, and it is based on a device I call the convergence zone. (Damasio 1989a: 45; my boldface.)

We'll come back to "binding codes" and the "convergence zone" later (Ch. **??**). Note here that the "fragments" correspond to lexical items, and the binding code corresponds to a syntactic structure. Thus, this brain-embodied language of thought is very compositional: The representation of an object consists of features plus a combinatorial arrangement of them.

1.9 Summary.

We began by considering the claim that there are two kinds of understanding: semantic and syntactic. The former is relational and is a correspondence between two domains. The latter is non-relational (or self-relational). In order to understand semantic understanding, we looked at a classical Tarskian semantic interpretation of a syntactic domain, whose lesson was that, in semantic understanding, one of the two domains must be antecedently understood. We then turned to data that supported the views that (1) there is a chain or "continuum" of syntactic and semantic domains, whose only difference is the role they play, and that (2) some domains can play *both* roles (the model muddle). Finally, we considered how mental models can be constructed computationally, and how they are fundamentally syntactic in nature. We now turn to a more detailed study of the second type of understanding: syntactic understanding.