UNDERSTANDING UNDERSTANDING: Semantics, Computation, and Cognition

William J. Rapaport

Department of Computer Science and Center for Cognitive Science
State University of New York at Buffalo, Buffalo, NY 14260
rapaport@cs.buffalo.edu, http://www.cs.buffalo.edu

DRAFT

17 July 1996

Contents

Abstract

What does it mean to understand language? John Searle once said: "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.?)." Elsewhere, I have argued "that (suitable) purely syntactic symbol-manipulation of a computational natural-language-understanding system's knowledge base suffices for it to understand natural language." The fundamental thesis of the present book is that understanding is recursive: "Semantic" understanding is a correspondence between two domains; a cognitive agent understands one of those domains in terms of an antecedently understood one. But how is that other domain understood? Recursively, in terms of yet another. But, since recursion needs a base case, there must be a domain that is not understood in terms of another. So, it must be understood in terms of itself. How? Syntactically! In syntactically understood domains, some elements are understood in terms of others. In the case of language, linguistic elements are understood in terms of non-linguistic ("conceptual") yet internal elements. Put briefly, bluntly, and a bit paradoxically, semantic understanding is syntactic understanding. Thus, any cognitive agent–human or computer–capable of syntax (symbol manipulation) is capable of understanding language. The purpose of this book is to present arguments for this position, and to investigate its implications. Subsequent chapters discuss: models and semantic theories (with critical evaluations of work by Arturo Rosenblueth and Norbert Wiener, Brian Cantwell Smith, and Marx W. Wartofsky); the nature of "syntactic semantics" (including the relevance of Antonio Damasio's cognitive neuroscientific theories); conceptual-role semantics (with critical evaluations of work by Jerry Fodor and Ernest Lepore, Gilbert Harman, David Lewis, Barry Loewer, William G. Lycan, Timothy C. Potts, and Wilfrid Sellars); the role of negotiation in interpreting communicative acts (including evaluations of theories by Jerome Bruner and Patrick Henry Winston); Hilary Putnam's and Jerry Fodor's views of methodological solipsism; implementation and its relationships with such metaphysical concepts as individuation, instantiation, exemplification, reduction, and supervenience (with a study of Jaegwon Kim's theories); John Searle's Chinese-Room Argument and its relevance to understanding Helen Keller (and vice versa); and Herbert Terrace's theory of naming as a fundamental linguistic ability unique to humans. Throughout, reference is made to an implemented computational theory of cognition: a computerized cognitive agent implemented in the SNePS knowledge-representation and reasoning system. SNePS is: symbolic (or "classical"; as opposed to connectionist), propositional (as opposed to being a taxonomic or "inheritance" hierarchy), and fully intensional (as opposed to (partly) extensional), with several types of interrelated inference and belief-revision mechanisms, sensing and effecting mechanisms, and the ability to make, reason about, and execute plans. This document is Technical Report 96-26 (Buffalo: SUNY Buffalo Department of Computer Science).

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

COMPUTATIONAL NATURAL-LANGUAGE UNDERSTANDING AND A COMPUTATIONAL MIND

1.1 UNDERSTANDING LANGUAGE.

> The question of how one understands the language one thinks in does seem to be a peculiar one. (Loewer 1982: 310.)

What does it mean to say that we have knowledge of the semantics of our language, and how do we come to have it? That is, what does it mean to say that we understand our language? And can a non-human—say, an ape or a computer—have such knowledge? What does it mean for *any* cognitive agent to understand language? Viewed from a very high distance, my answer is this:

> "Semantic" understanding is a correspondence between two domains; a cognitive agent understands one of those domains in terms of the other.

But if a domain, A, is to be understood in terms of another, B, how, then, does the agent understand domain B? The simplest answer, I believe, is, in good recursive fashion, that B is also understood in terms of a domain, namely—since good recursion needs a base case—itself! But how can something be understood in terms of itself? The answer is: syntactically. Put briefly, bluntly, and a bit paradoxically, my thesis is, at bottom, this:

Semantic understanding is syntactic understanding.

And any cognitive agent, human or computer, that is capable of syntax—of symbol manipulation—is, thus, capable of understanding language.

Clearly, we need to zoom in from this broad overview and focus on the details. Along the way, we shall consider a seeming grab-bag of topics: computer implementations, the Chinese-Room

Argument, conceptual-role semantics, methodological solipsism, natural-language understanding, language understanding by apes, computational lexicography, newspaper photographs with captions, and even Helen Keller.

Let me begin with a brief review of some of the claims I made in "Syntactic Semantics" (Rapaport 1988b) about natural-language understanding by computers. I beg the reader's indulgence: I will be making many controversial claims here, based on many implicit assumptions, that I will not justify or make explicit until later. For now, please consider what follows to be part of my thesis as seen from a slightly lower, though still quite high, vantage point. I'm still merely trying to show you the lay of the land.

1.2 LANGUAGE AND MIND.

> ... the crucial feature of cognitive practice [is] ... the ability to make representations. (Wartofsky 1979: xiii.)

1.2.1 Computers, Programs, and Processes.

Strictly speaking, neither computers nor programs can understand natural language. I do not say this in the uninteresting sense that no present-day computers or AI programs can understand natural language. I readily admit that some suitably-programmed computers can process quite a lot of natural language, though none can do it (yet) to at least the degree that would be needed to pass a Turing Test. Rather, if a suitably programmed computer is ever to pass a Turing Test for natural-language understanding, what will understand natural language will be neither the mere physical computer (the hardware) nor the static, "inert" (Smith 1987: 15, 17), textual program (the software), but the dynamic, "active" (Smith 1987: 17) behavioral *process*—the program being executed by the computer. As Andrew S. Tanenbaum (1976: 12) puts it (cf. Smith 1987, sect. 5):

> A **process** ... is basically a program in execution. It is an active entity, capable of causing events to happen. A process (i.e., a running program) may draw Chinese characters on the pen-and-ink plotter, or it may play chess with a human being sitting at a time-sharing terminal ... . A process is in contrast to a program, which is a passive entity. A program lying on someone's desk cannot do anything by itself.

1.2.2 The Real Thing.

Such a successful natural-language–understanding process will be an example of "strong AI" in two senses of that phrase. First, the process will probably be "psychologically valid"; i.e., the underlying algorithm will probably be very similar (if not identical) to the one we use. Second, natural-language understanding is at least necessary, and possibly sufficient, for passing the Turing Test. Thus, anything that passes the Turing Test does understand natural language. But such a process will pass the Turing test. So, such a process will do more than merely simulate natural-language understanding; it will *really* understand natural language.

Or so I claim. Any process that understands natural language will have to have a number of features. I introduce the main ones in the next two sections.

1.2.3 Robustness.

A natural-language–understanding process must be "open-ended" or "robust": It will have to be able to deal with what might be called "improvisational audience-participation discourse".

First, some "canned" patterns of conversation will be needed, as Schank, Minsky, and others have urged with their theories of "scripts", "frames", etc. (Schank & Rieger 1974, Schank 1975, Minsky 1975, Schank & Riesbeck 1981). But it cannot rely solely on these. Just as we can use language in arbitrary and unforeseen circumstances, so must the computational process be able to improvise.

Second, the process must be capable of conversing with an interlocutor. Monologues and talking to (or writing essays for) oneself are fine as far as they go; but a language-using entity must be able to converse with an interlocutor, else it would not pass the Turing Test. Such interaction provides feedback, allowing the two natural-language understanding systems—the two interlocutors—to reach mutual understanding (to bring their "knowledge bases" into "alignment"). It also provides causal links with the outside world. But I will leave open for now such questions as what kind of links these are, what their role is in semantics, and how important they are. (To tantalize you, I think they play virtually no role and are irrelevant!)

Finally, the process must be able to do more than understand isolated sentences (so beloved of many philosophers and linguists). It must be able to understand *sequences* of sentences that form a coherent discourse. What it understands at any point in a discourse will be a function partly of what it understood before. As Erwin Segal et al. (1991: 32) put it in the context of understanding narrative text,

> A new sentence in the text is interpreted in terms of an ongoing construction of an integrated component of the narrative's meaning. Unless specifically marked, the new meaning is incorporated into, and regarded as continuous with, the current ongoing construction.

The same holds for understanding language in general. (On potential *dis*continuities, due to shifts between "objective" and "subjective" contexts, see Wiebe & Rapaport 1988; Wiebe 1990, 1991, 1994; Duchan et al. 1995.)

1.2.4 Natural-Language Competencies.

There are many things a natural-language–understanding process needs to be able to do.

1. It must understand virtually all input—virtually all that it "hears" or "reads"—whether grammatical or not; after all, *we* do.

2. It must remember what it knew or heard before, as well as ...

3. ... what it learns in the course of a conversation.

4. It must be able to perform inference on what it hears and what it knows; ...

5. ... revise its knowledge or beliefs, as needed; ...

6. ... and remember what, that, how, and why it inferred.

7. It must be able to make plans ...

8. ... and to understand plans: Among the plans that it must be able to make and to understand, it must be able to plan speech acts, so that it can *generate* language ...

9. ... to answer questions, ...

10. ... to ask questions, ...

11. ... and to initiate conversation. Thus, by the way, it would not be merely a natural-language–*understanding* process, but a natural-language–understanding and –*generation* process—what Stuart C. Shapiro and I have elsewhere called natural-language *competence* (Shapiro & Rapaport 1991).

12. And it must be able to understand the speech-act plans of its interlocutors, in order to understand why speakers say what they do.

13. This, in turn, requires the process to have (or to construct, on the fly) a "user model"—a theory of what the interlocutor believes.

14. Last on this list (though no doubt more is needed), the cognitive agent who understands natural language must, as part of its ability to use and understand natural language, be able to learn via language—to learn about non-linguistic things (the external world, others' ideas), ...

15. ... and to learn about language, including its own language (e.g., it must be able to learn its own language from scratch, as we do from infancy, as well as consciously learn the syntax and semantics of its language, as we do (or should) in school).

1.2.5 Mind.

To do all of this, a cognitive agent who understands natural language must have a "mind", or what AI researchers call a 'knowledge base'. It will, initially, contain what might be called "innate ideas"—anything in the knowledge base before any language use begins. And it will come to contain beliefs resulting from perception, from conversation, and from inference. Among these will be internal representations of external objects.

For convenience as well as for the sake of perspicuousness, we will think of the knowledge base or mind as a propositional semantic network, whose nodes represent individual concepts, properties and relations, and propositions, and whose connecting arcs structure atomic concepts into molecular ones (which include structured individuals, propositions, and rules). The specific semantic-network theory we will use throughout is the SNePS knowledge representation and reasoning system (see §1.3), but the reader is invited to think in terms of other knowledge representation and reasoning systems as well, such as (especially) Discourse Representation Theory (Kamp 1983, 1984, 1985, 1988; Asher 1986; Kamp & Reyle 1993), the KL-ONE family (Brachman & Schmolze 1985, Woods & Schmolze 1992), Conceptual Dependency (Schank & Rieger 1974, Schank 1975, Schank & Riesbeck 1981, Hardt 1992, Lytinen 1992), or Conceptual Graphs (Sowa 1984, 1992).

1.2.6 Syntax Suffices.

> Philosophy must be done in the first person, for the first person. (Hector–Neri Castañeda, in conversation, 1984)

Meaning will be, *inter alia*, relations among these internal representations of external objects, on the one hand, and other internal symbols of the language of thought, on the other. A cognitive agent, C, who has natural-language competence as described so far, understands the natural-language output of another such cognitive agent, O, "by building and manipulating the symbols of an internal model (an interpretation) of [O's] output considered as a formal system. [C]'s internal model would be a knowledge-representation and reasoning system that manipulates symbols" (Rapaport 1988b: 104). This is the way in which C's semantic understanding of O is a *syntactic* enterprise.

There are two semantic points of view that must be carefully distinguished. The *external* point of view is C's understanding of O. The *internal* point of view is C's understanding of itself. And there are two ways of viewing the external point of view: The "third-person" way, in which we, as external observers, describe C's understanding of O, and the "first-person" way, in which C understands its own understanding of O (see Ch. 6). Traditional referential semantics is largely irrelevant to the latter, primarily because external objects are—*can*—only be dealt with via internal representations of them. It is the latter—first-person, internal understanding—that I seek to understand and that, I believe, can only be understood syntactically.

The first three quotations that open this book, from Hohenberger, Bush, and Searle, outline my argument: (1) All of our beliefs are from the first-person point of view. Therefore, (2) they are internal. Consequently, (3) they form a single, *syntactic*, system. Following the admonition in Shapiro's opening quotation, the rest of this book is an investigation into what kind of sense this makes.

1.3 A COMPUTATIONAL MIND.

As noted, to help fix our ideas, it will be useful to talk in terms of a specific knowledge-representation and reasoning system. The one I will use is the SNePS *S*emantic *Ne*twork *P*rocessing *S*ystem, developed by Stuart C. Shapiro and colleagues (primarily in the SNePS Research Group at State University of New York at Buffalo; cf. Shapiro 1979; Shapiro & Rapaport 1987, 1992, 1995). *As a knowledge-representation system,* SNePS has the following features: It is:

1. symbolic (or "classical"; as opposed to connectionist),
2. propositional (as opposed to being a taxonomic or "inheritance" hierarchy),
3. fully intensional (as opposed to (partly) extensional).

As a reasoning system, it has several types of interrelated inference mechanisms:

1. "node-based" (or "conscious") inference,
2. "path-based" inference (generalized inheritance or "subconscious" inference),
3. "default" reasoning, and

4. belief revision.

Finally, it has certain *sensing and effecting mechanisms*, namely:

1. perception,[1]

2. natural-language competence, and

3. the ability to make, reason about, and carry out plans.

Such, at least, is SNePS in principle. Various implementations of it have more or less of these capabilities, but I will assume the ideal, full system and sprinkle examples from the implementations throughout.

There is no loss of generality in focussing on such a system. Its symbolic nature begs no questions vis-à-vis connectionism. A connectionist system that passed the Turing Test would make my points about the syntactic nature of understanding equally well. For a connectionist system is just as computational—just as syntactic—as a classical symbolic system (Rapaport 1993a).

That SNePS is propositional rather than taxonomic merely means that it is capable of representing propositions as well as taxonomic hierarchical relationships among individuals and classes. Actually, it represents such relationships propositionally, and the automatic inheritance feature of such systems is generalized to path-based inference in SNePS. Some knowledge-representation and reasoning systems are event-based or "situation"-based; both events and situations can also be represented in SNePS.

But SNePS is intensional, and therein lies a story. 'Intensional' is an overworked term. In the present context, it means this: To be able to model the mind of a cognitive agent, a knowledge-representation and reasoning system must be able to represent and reason about intensional objects, by which I mean objects not substitutible in intensional contexts (such as the morning star and the evening star), indeterminate or incomplete objects (such as fictional objects), non-existent objects (such as a golden mountain), impossible objects (such as a round square), distinct but coextensional objects of thought (such as the sum of 2 and 2 and the sum of 3 and 1), and so on (cf. Rapaport 1991a, Shapiro & Rapaport 1991). Why bother with these? Because we think and talk about them, and therefore so must any entity that uses natural language. Non-intensional knowledge-representation systems just won't do (*pace* Wilensky 1991).

We are going to use SNePS to model the mind of a cognitive agent. The agent whose mind is thus implemented in SNePS is named 'Cassie'. If we need to model a second such agent, we'll call him 'Oscar'.[2] If Cassie passes the Turing Test, then she *is* intelligent and *has* (or perhaps *is*) a mind. (Or so I claim.) Her mind consists of SNePS nodes and arcs. SNePS, that is, is her language of thought (in the sense of Fodor 1975). If she is implemented on, say, a Sun workstation, then we might also say that she has a "brain" whose components are the "switch-settings"—the contents of the registers—in the Sun that implements the nodes and arcs of her mind.

[1]Though only to a limited degree at present; cf. Srihari & Rapaport 1989, 1990; Srihari 1991ab; Lammens 1994; Lammens et al. 1995; Hexmoor et al. 1993abc; Hexmoor 1995.

[2]Cassie is the *C*ognitive *A*gent of the *S*NePS *S*ystem—an *I*ntelligent *E*ntity. Oscar is the *O*ther *S*NePS *C*ognitive *A*gent *R*epresentation. Cassie was first introduced in the 1985 ancestor of Shapiro & Rapaport 1987. Oscar was first introduced in Rapaport, Shapiro, & Wiebe 1986.

To be more accurate about what can be represented in SNePS—better, what can be represented in SNePS/Cassie; better yet, what Cassie can represent—we will say that Cassie can represent—or think about—objects (whether existing or not), properties and relations, propositions, events, situations, etc. Thus, all of the things represented in SNePS when it is being used to model Cassie's mind are objects of Cassie's thoughts (i.e., Meinongian objects of Cassie's mental acts); they are, thus, inten*t*ional—hence inten*s*ional—objects. They are not extensional objects in the external world, though, of course, they may bear some relationships to such external objects. (See Shapiro & Rapaport 1991.)

I shall not rehearse here the arguments I and others have made elsewhere for these claims about SNePS and Cassie. I will, however, provide examples of SNePS networks in the chapters that follow. (For further examples and argumentation, the reader is urged to consult Maida & Shapiro 1982; Shapiro & Rapaport 1987, 1991, 1992, 1995; Rapaport 1988b, 1991a; and Rapaport & Shapiro 1995. These should be thought of as forming an integral part of the present essay.)

Let us move on, now formulating our questions in terms of Cassie: Can we say that Cassie understands English? If so, how does she? John Searle, of course, would say that she doesn't. I say that she does and that she does so by manipulating the symbols of her language of thought, viz., SNePS. Let's turn now to these issues.

Chapter 2

SEMANTICS AS CORRESPONDENCE.

2.1 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 2.1.)

[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.

Figure 2.1: The relation of syntax to semantics.

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.

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 9.)

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.

---

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

But note that they are, nonetheless, relationships *among linguistic, that is, syntactic, items*. Hence, on our terms, they, too, are "syntactic", not "semantic" (cf. §3.2.1, 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.

2.2 TARSKIAN SEMANTICS.

2.2.1 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 $\ulcorner\neg\varphi\urcorner, \ulcorner(\varphi \vee \psi)\urcorner, \ulcorner\forall v[\varphi]\urcorner$ 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 (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.

2.2.2 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 §2.7).

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 **D**? Whichever you want, or, if we spell out $\mathcal{L}$ and **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 **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 **D** that are the interpretations of the t_i; and the result will be an element of **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 **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, **M** is a virtual duplicate or mirror image of $\mathcal{L}$. (Indeed, **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 **M**; in that case, I would be a *partial* function. Such is the case when $\mathcal{L}$ is English and **M** is the world ('unicorn' is English, but unicorns don't exist), though if we "enlarge" or "extend" **M** in some way, for example, if we take **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 **M** than in $\mathcal{L}$; that is, there are elements of **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 **M** (and a different I), we could give the symbols of $\mathcal{L}$ entirely distinct meanings. Worse, the two **M**s need not be isomorphic. (This can happen in at least two ways. First, the cardinalities of the two **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 **M**. To do that, I need to show you **D**: **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). **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 **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 **R**, of course, is part of **M**); that is, R means the deducibility relation on wffs of propositional logic. So, the elements of $\mathcal{R}$ are rules of inference:

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

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

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)

$I(\mathcal{R}11)$ 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.

2.3 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 2.1 and 2.2.) 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*.

   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.

	role of the syntactic domain	role of the semantic domain
1.	a formal language $\mathcal{L}$	a model **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 2.2)	
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 2.1: Syntactic and semantic domains.

$$\text{narrative} \rightarrow \text{play} \rightarrow \begin{cases} \text{opera} \rightarrow \text{ballet}_1 \rightarrow \text{film}_1 \rightarrow \text{novelization}_1 \\ \text{ballet}_2 \\ \text{film}_2 \rightarrow \text{novelization}_2 \rightarrow \text{film}_3 \\ \text{symphony} \rightarrow \text{performances} \end{cases}$$

Table 2.2: 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.

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 §1.2.5. 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*).

2.4 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 3 and 4), we need to look at semantic correspondences more carefully.

2.5 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 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 (§2.7.3). 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".

Figure 2.2: How to make the semantic domain fit the syntactic domain.

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 2.2). 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).

2.6 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.

2.6.1 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. 7). 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 (§2.7.1).

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 (§2.4). 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.

[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.

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 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 Os 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).)

2.6.2 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 §2.3. 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.

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

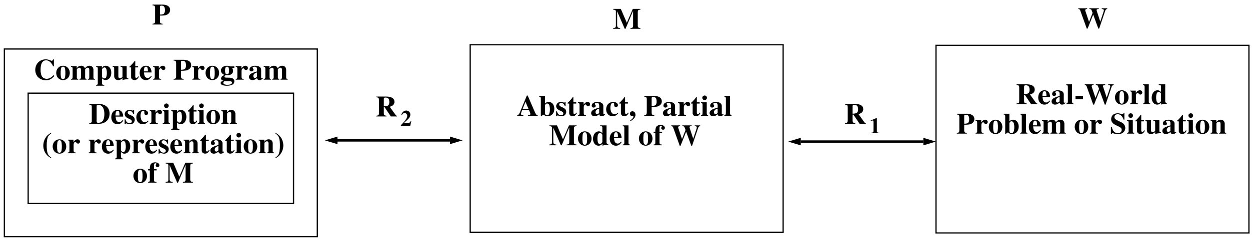


Figure 2.3: Smith's version of the "model muddle".

2.7 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').

2.7.1 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 2.3 (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

> 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 §§3.2.2.2, 5.1, 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

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

could not pass the Turing Test; cf. §1.2.3.)

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!

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 (§2.8.2, and Chs. 3 and 5).

2.7.2 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

[8] Cf. Helen Keller's labels; see Ch. 9.

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 2.3. 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 *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).

2.7.3 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 (§2.6.2).

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. §4.2, 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 (§2.7.1), 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. §2.3, 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 §2.3. 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 2.4).

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 2.5. 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 2.6, a version of Figure 2.4 (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

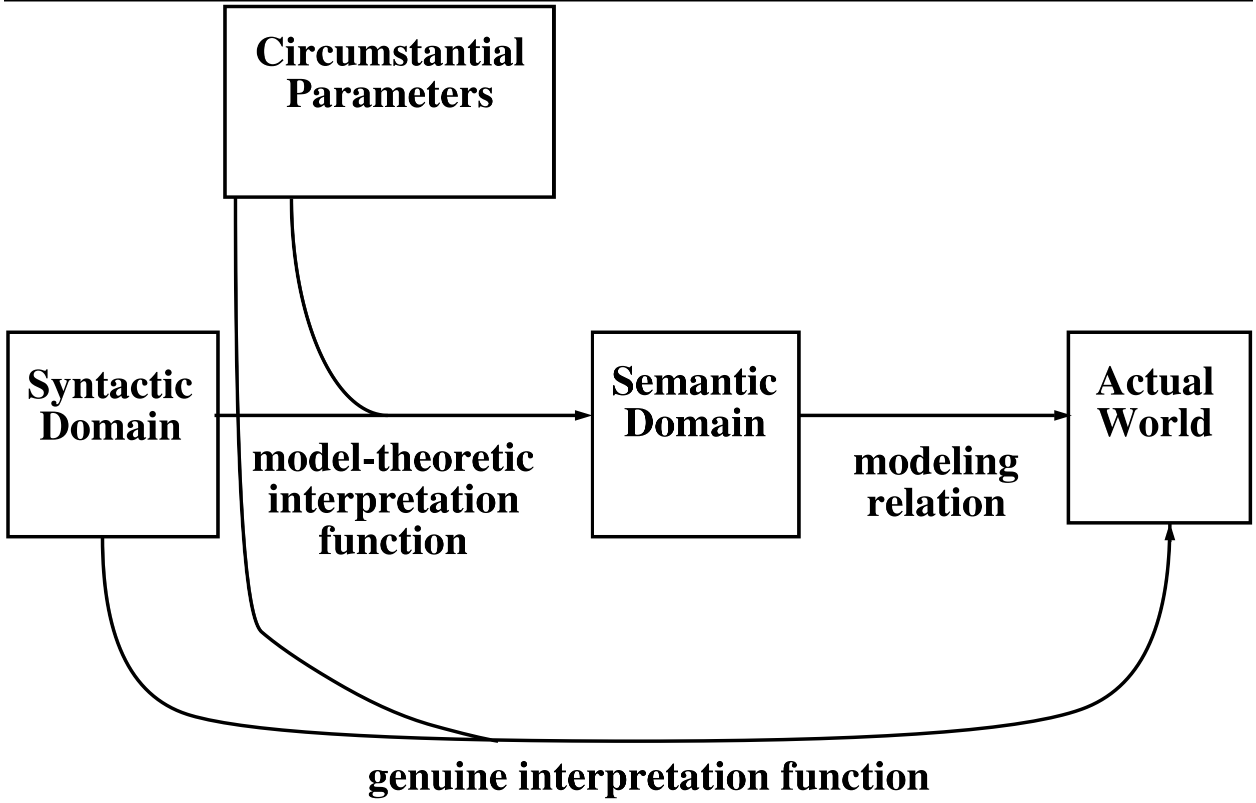


Figure 2.4: Smith's commutative semantic diagram.

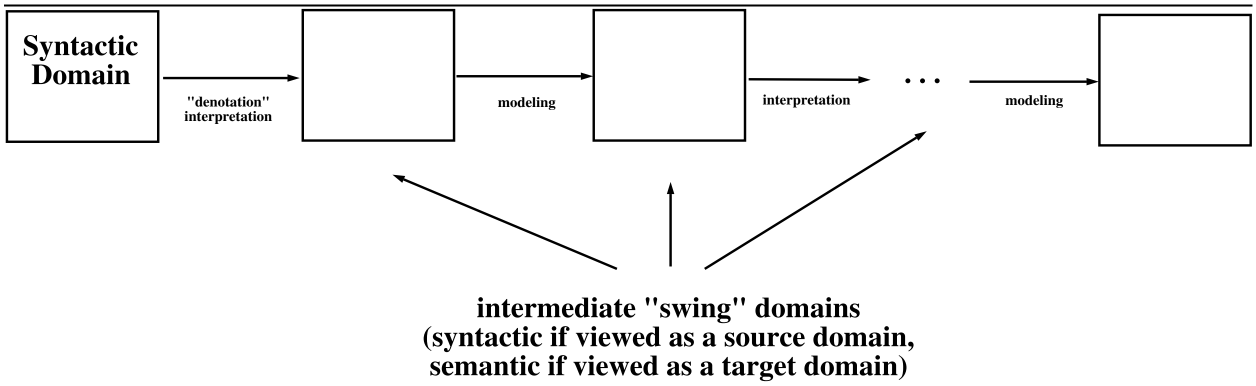


Figure 2.5: Smith's correspondence continuum.

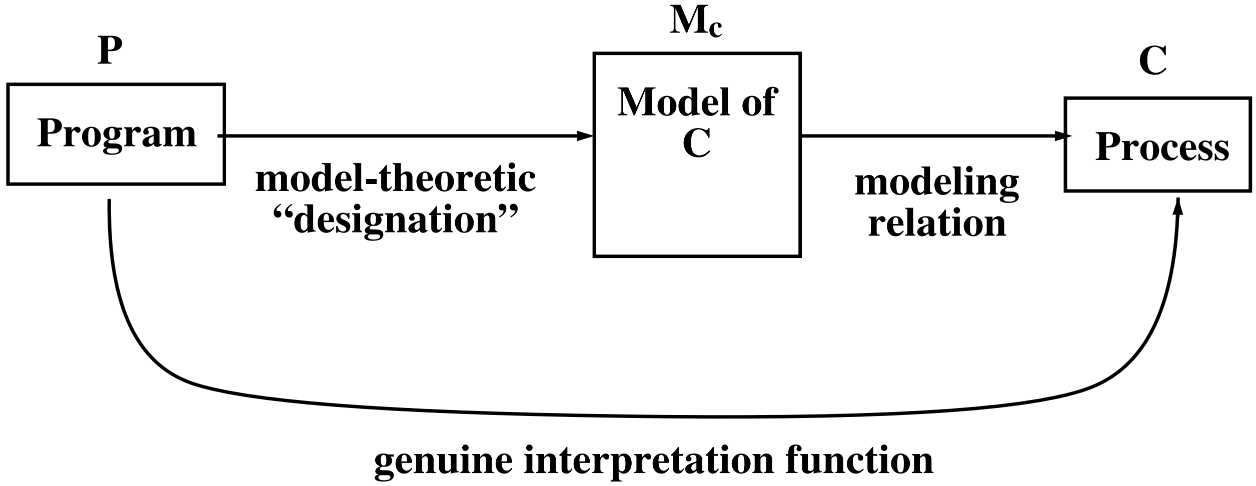


Figure 2.6: Smith's view of programs and processes.

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 2.7 (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". 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 2.8. 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

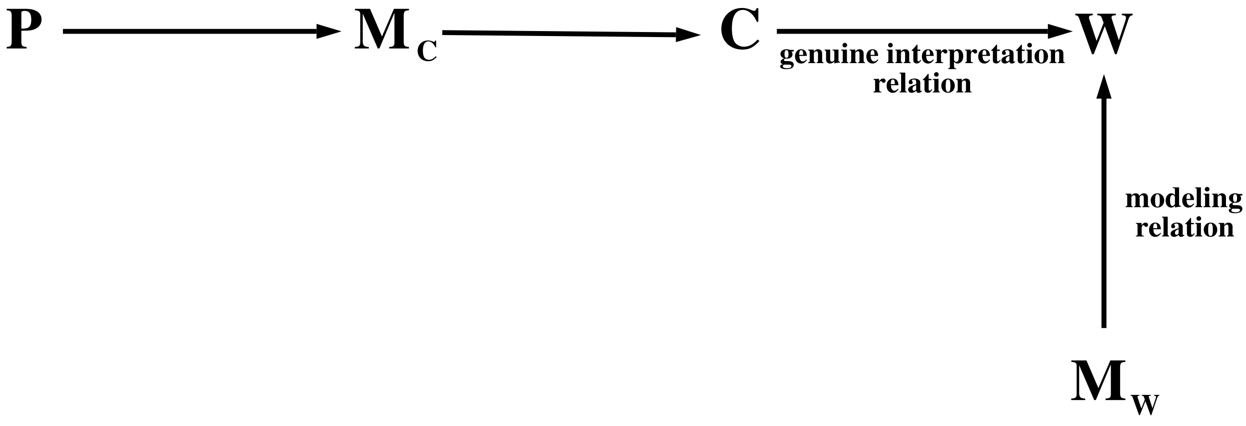


Figure 2.7: Smith's view of programs and processes, elaborated.

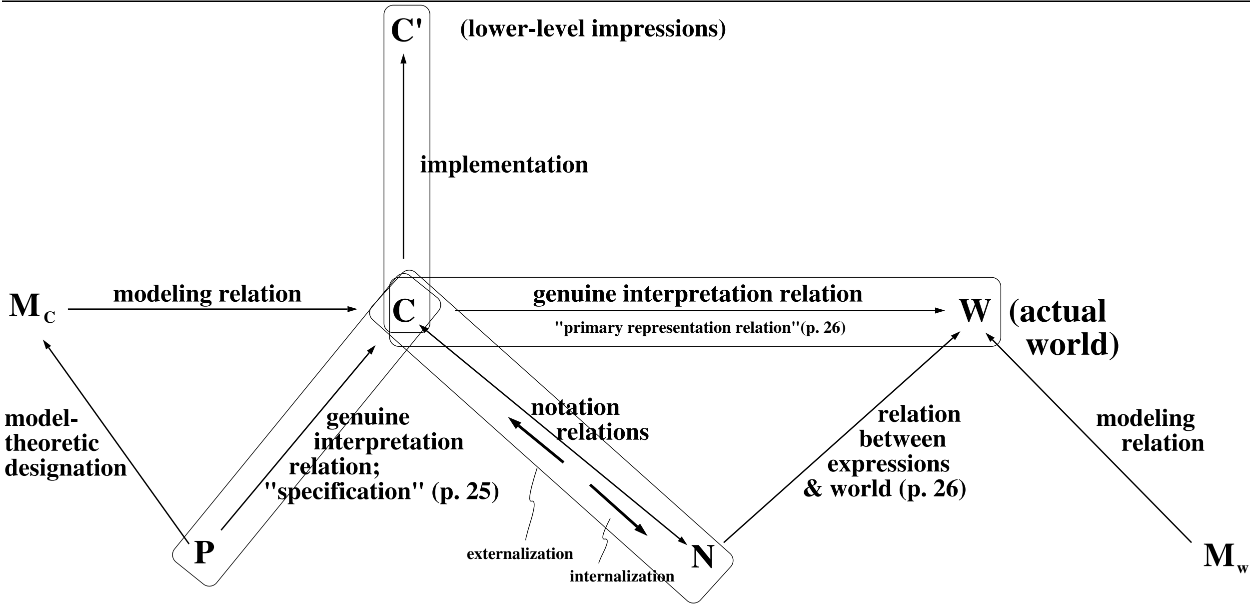


Figure 2.8: Smith's full picture.

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 7).

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 2.9, 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.

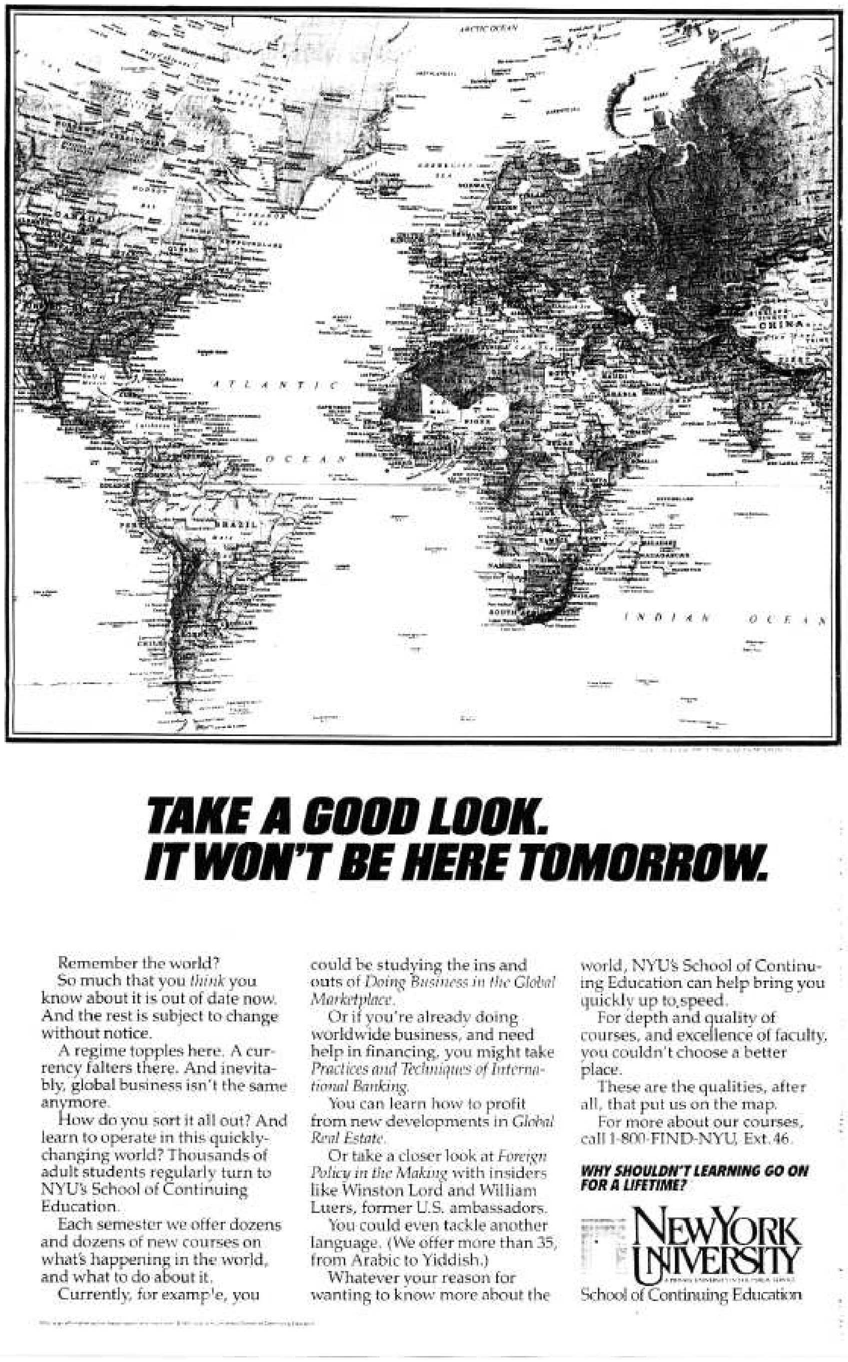


Figure 2.9: Ad for NYU, *New York Times* (20 August 1991: D5). Is this a map? A photo of a map? A reproduction of a photo of a map?

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 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.

2.7.4 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 ($\mathrm{M_W}$), and a linguistic description (in some program) of the model ($\mathrm{D_{M_W}}$), what we have in Cassie's case is that her mental model of the world is simultaneously $\mathrm{M_W}$ and $\mathrm{D_{M_W}}$. 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.

2.8 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.

2.8.1 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 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.

[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.

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, §5.4.

2.8.2 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 1, 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]

[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

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. §1.2.4, 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 §9.2).

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 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 §3.2.2.1. 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,

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.)

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. 4).

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 2.10. (Linearly abbreviated: `M2` = `B1` is named 'Lucy'; `M4` = `B1` is rich). Thus, Cassie might think to herself something like: "My thought that $\texttt{LUCY}\overset{\texttt{M2!}}{\wedge}\texttt{B1}\overset{\texttt{M4!}}{\wedge}\texttt{RICH}$ 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 (Maida & Shapiro 1982, Shapiro & Rapaport 1987), this network should—and could—be re-used), and (4) a biconditional rule asserting the equivalence (see Figure 2.11 for a possible version). Thus, the best

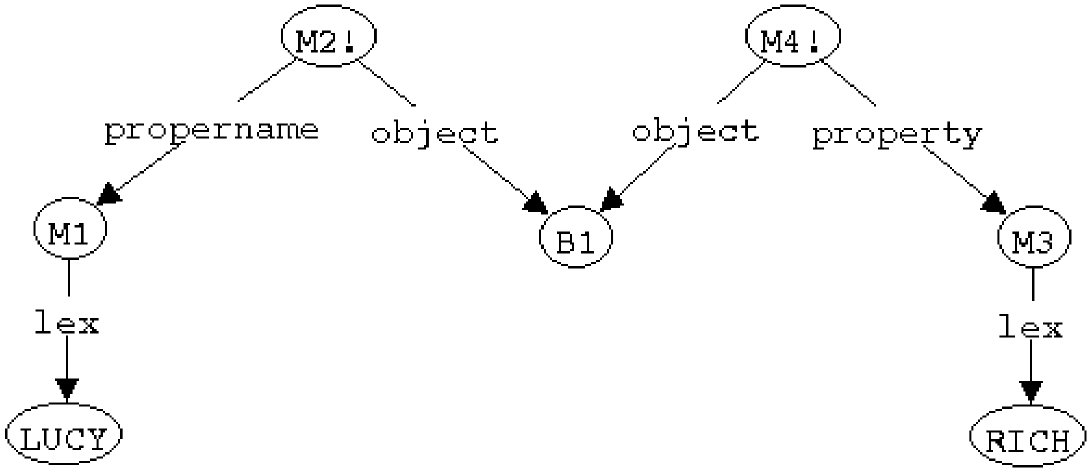


Figure 2.10: Cassie's belief that Lucy is rich.

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 2.11) 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

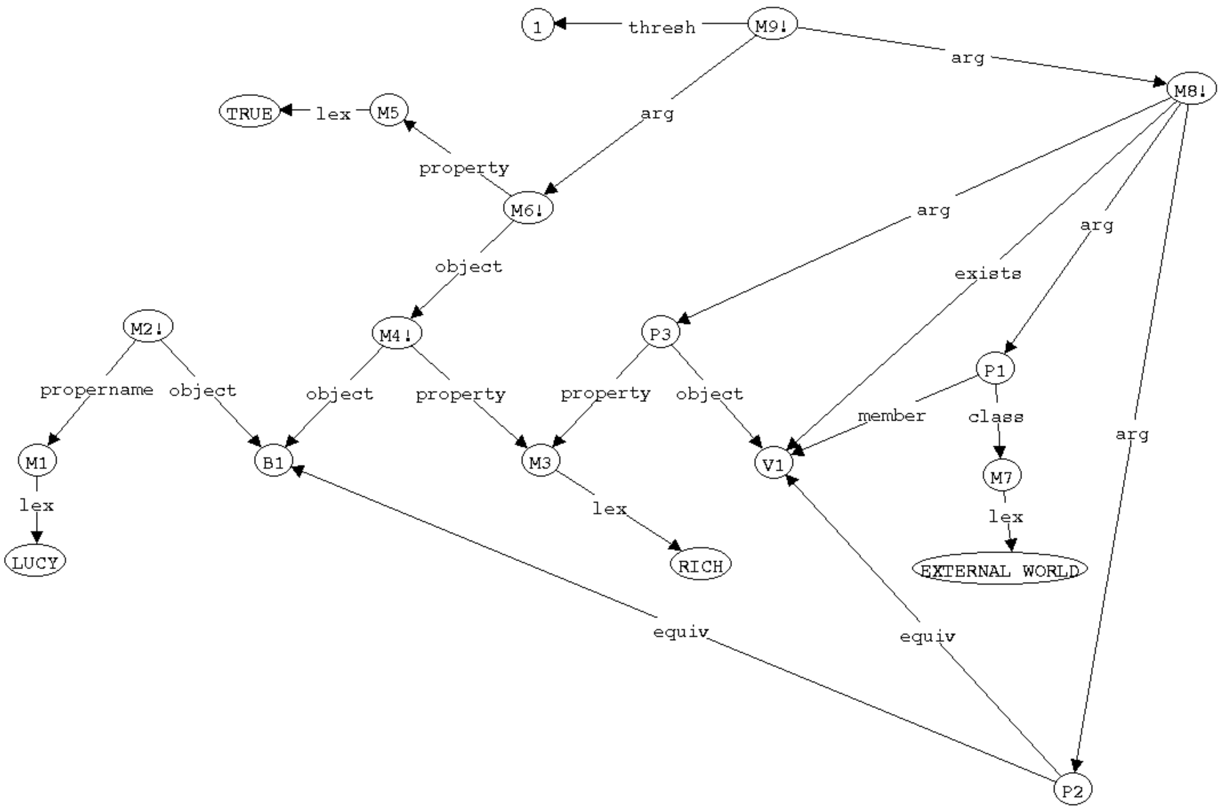


Figure 2.11: A biconditional rule (`M9`) asserting the equivalence of `M6` = that Lucy is rich is true, and `M8` = something in the external world is Lucy and is rich. More fully:

`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`.)

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 *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 2.12). 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 $den(t, w) = f(t) = \texttt{M1}$. If so, what's the relation between `M1` and `B1`? If $t = \texttt{M1}$, on the other hand, then $f(t) = \texttt{B1}$; 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 Jeannette Neal (1981, 1985; Neal & Shapiro 1984, 1985, 1987). (On this view, $t = \texttt{LUCY}$ (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 2.13. Now, her *understanding*—her semantic interpretation—of that sentence is the belief shown in Figure 2.10. Then:

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

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

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

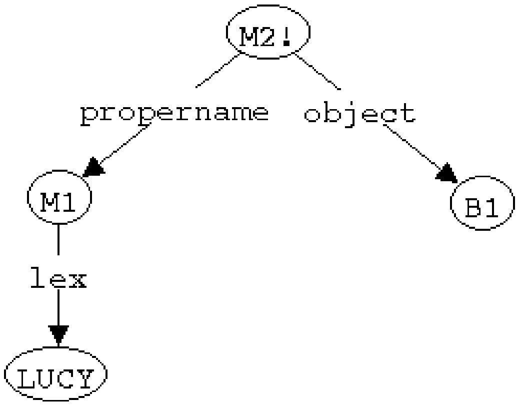


Figure 2.12: Cassie's belief that someone (`B1`) is named 'Lucy'.

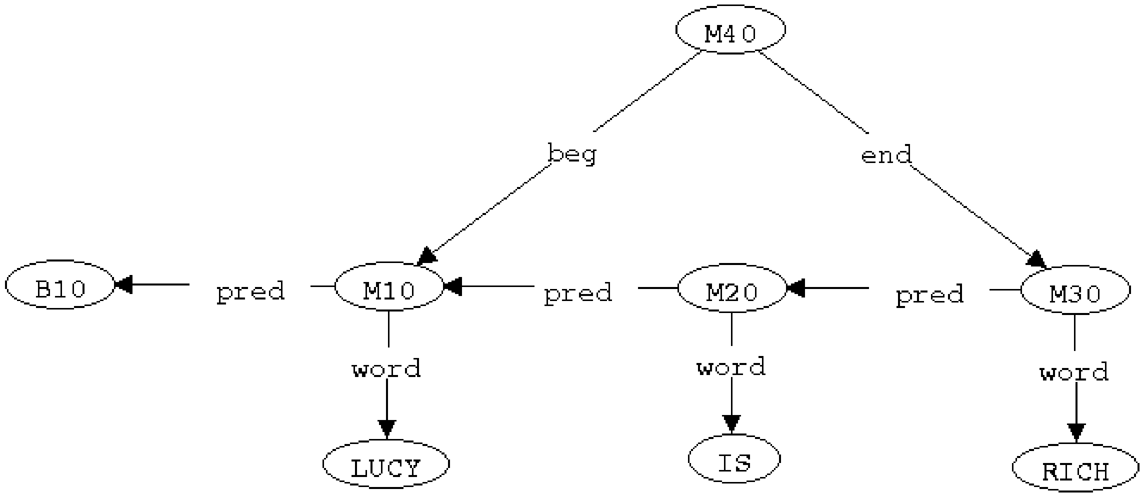


Figure 2.13: Cassie's representation of the *sentence* 'Lucy is rich' (roughly, `M40` = the sequence of words beginning with 'Lucy' and ending with 'rich'; after Neal & Shapiro 1987: 63).

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. 2.11) 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. 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 2.14. 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

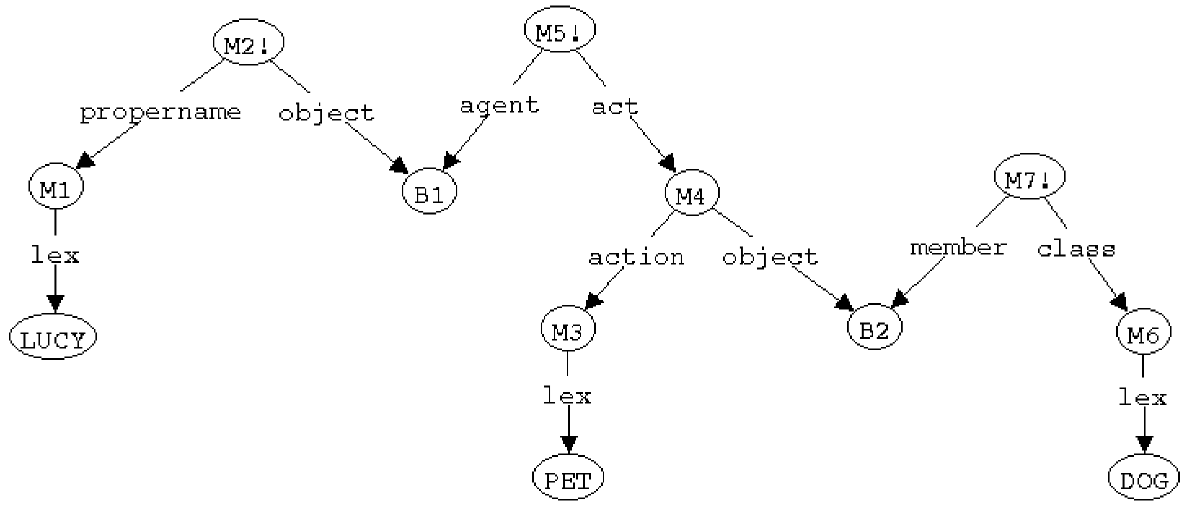


Figure 2.14: Cassie's belief that Lucy pets a dog:
M2 = B1 is named 'Lucy';
M7 = B2 is a dog;
M5 = B1 pets B2.

the following (sub)structure: It consists of an 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 2.15), 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, expressed by 'coat'.

Syed Ali's ANALOG system (Ali 1994, 1995; Ali & Shapiro 1993; see Figure 2.16) 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

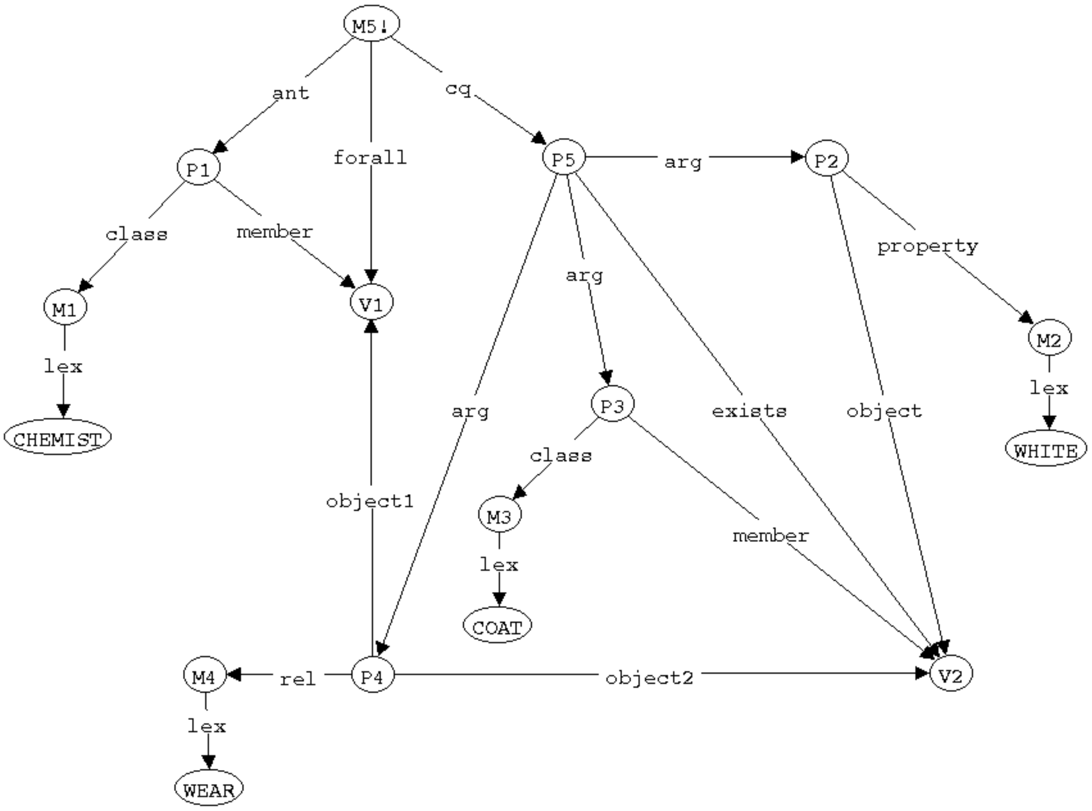


Figure 2.15: Cassie's belief that all chemists wear white coats:
M5 = $\forall$V1[P1 $\rightarrow$ P5];
P1 = V1 is a chemist;
P5 = $\exists$V2[P2 & P3 & P4];
P2 = V2 is white;
P3 = V2 is a coat;
P4 = V1 wears V2.

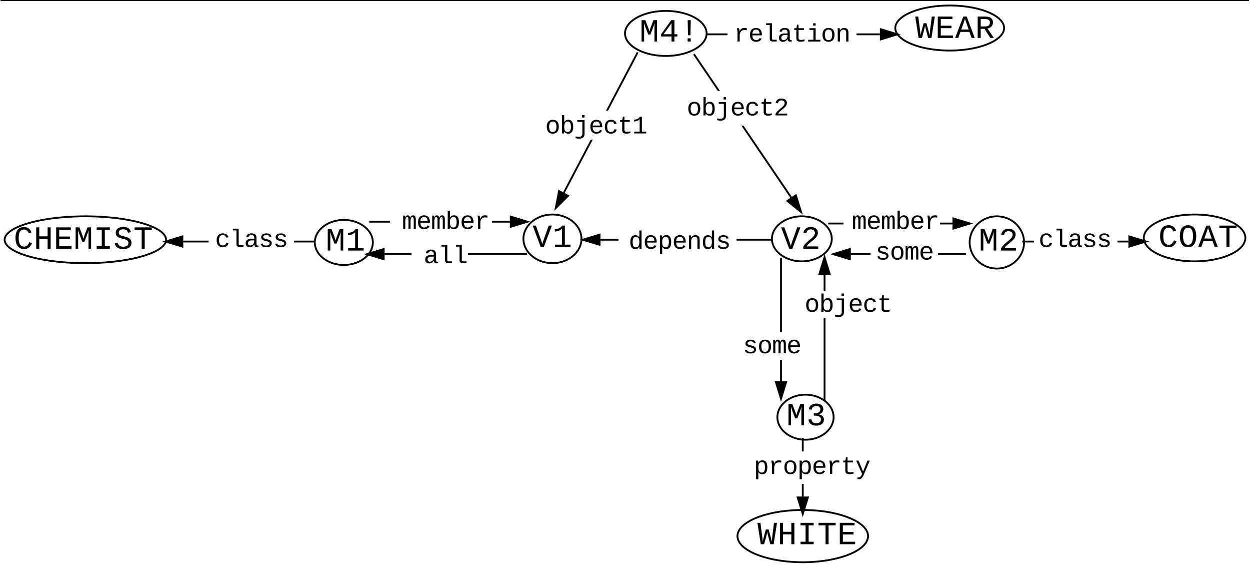


Figure 2.16: ANALOG representation of 'All chemists wear white coats':
M4 = V1 wears V2.
M1 = V1 is an arbitrary member of the class of chemists.
M2 = V2 is a member, depending on V1, of the class of coats and which is white.

(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.

2.8.3 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. 3). 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.

2.9 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.

Chapter 3

SEMANTICS AS SYNTAX.

> ... the correspondence continuum challenges the clear difference between "syntactic" and "semantic" analyses of representational formalisms .... no simple "syntactic/semantic" distinction gets at a natural joint in the underlying subject matter. (Smith 1987: 38.)

3.1 SUMMARY AND INTRODUCTION.

Here is the story so far: To understand language is to construct a semantic interpretation—a model—of the language. In fact, *most* understanding is like this: We normally understand something by modeling it and then determining correspondences between the two domains. In some cases, we are lucky: We can, as it were, keep an eye on each domain, merging the images in our mind's eye or, to switch metaphors, "co-activating" the two domains (cf. Mayes 96: 111). In other cases, notably when one of the domains is the external world, we are not so lucky—Smith's Gap cannot be crossed—and so we can understand that domain *only* in terms of the model. Lucky or not, we understand one thing in terms of another by modeling that which is to be understood (that is, the syntactic domain) in that which we already understand (that is, the semantic domain). For this to yield understanding of the syntactic domain, the model must be antecedently understood.

Antecedent understanding has a long and distinguished history: Had none of the three languages on the Rosetta Stone been antecedently understood, we would not have been able to understand the others. (As it happened, only one was antecedently understood: the Greek. And the Greek text stated that all three texts said the same thing (an interesting use of an essential indexical; cf. Perry 1979). Hence, with a little bit of help from Champollion's antecedent understanding of Coptic ("a direct descendent of ancient Egyptian"), the demotic and hieroglyphic versions were able to be understood. Cf. Quirke & Andrews 1988: 3.)

Imagine Helen Keller after the well house, returning home to her antecedently familiar surroundings: "... every object which I touched seemed to quiver with life. That was because I saw everything with the stronger, new sight that had come to me" (Keller 1905: 36). But her "new sight" would not have been as effective had she not antecedently "understood" her surroundings.

Even so humble an expression as 'The lamp is near the radio', which takes the lamp as figure to the radio's ground, locates the lamp in terms of the antecedently known location of the radio. It is not informationally equivalent to 'the radio is near the lamp' if you're seeking the lamp. An antecedently understood semantic domain serves as ground for the syntactic domain as figure (cf. Talmy 1978). In the lucky case, the figure and ground are simply highlighted portions of a single domain, not two separate domains. We will see later (§3.2.2.2) how minds can always be lucky (we have seen it before in the relation of linguistic and non-linguistic nodes, §2.8.2).

But how is the antecedently understood domain antecedently understood? In the base case of our recursive understanding of understanding, a domain must be understood in terms of *itself*, that is, syntactically. We have seen (§2.3, example 27) that Smith considers the *syntactic* mapping Ψ from internal elements to other internal elements to be "*semantical*" (Smith 1982: 10). But we will also investigate other options.

3.2 SYNTACTIC UNDERSTANDING.

3.2.1 Familiarity Breeds Comprehension.

What is this syntactic mode of understanding? What does it mean to "get used to" something? In some sense, it should be obvious. Consider the following anecdote:

> In today's chess, only the familiarly shaped Staunton pieces are used. ... [One] reason is the unfamiliarity, to chess players, of other than Staunton pieces. ... [In Reykjavik, in 1973, two grandmasters] started to play [with a non-Staunton set], and the conversation ran something like:
>
> "What are you doing? That's a pawn."
> "Oh. I thought it was a bishop."
> "Wait! Maybe it is a bishop."
> "No, maybe it really is a pawn."
>
> Whereupon the two grandmasters decided to play without the board. They looked at each other and this time the conversation ran:
>
> "D5"
> "C4"
> "E6"
> "Oh, you're trying *that* on me, are you? Knight C3."
>
> And they went along that way until they finished their game. (Schonberg 1990: 38–39.)[1]

In a game played with Staunton pieces, the players are "used to" the pieces. Even in a game played with no physical pieces at all, the players are "used to" the symbolic notation for the pieces. But in a game played with non-Staunton pieces, clearly they are not.

[1]Cf. a similar conversation, in a language of "nerve states", in Eco 1988.

Or consider this comment by my high-school English teacher about "the way I judge poetry. Having read the best poetry for sixty years, I KNOW what's good" (Spencer Brown, personal communication, 1988). He has "gotten used to" poetry and so can judge it.

Suppose, as we did in the previous chapter, that the semantic relation is (merely) a correspondence relation. Suppose, further, as we did in §2.2.2, that it is a homomorphism—that is, a structure-preserving (or compositional) function mapping the syntactic domain into the semantic domain. Now, to understand something in terms of itself would then be to take the syntactic domain as its own semantic domain, treating the homomorphism as an *automorphism* (that is, a *self*-homomorphism), mapping the syntactic domain into itself. Such an automorphism would be a relation *among* the symbols of the syntactic domain, hence a classically syntactic relation. Yet it would also be a *semantic* relation—because it is a correspondence between "two" domains (better: between two roles played by the same domain). Indeed, Chang and Keisler's very first example of a semantic model in their *Model Theory* (1973) is such a mapping. The syntactic domain, now considered as its own *semantic* domain, is syntactic in the classical sense: It is a domain of symbols related in certain ways (by the automorphism). Thus, it is syntactic twice over, so to speak: once by way of its own, purely syntactic, features, and once by way of the semantic automorphism. (Recall the way some linguists do semantics; cf. §2.1.)

One must, I suppose, be careful not to get carried away: "In a lecture, the professor wrote on the blackboard: $\lim_{x \to 0+} \frac{8}{x} = \infty$ (infinity). On the subsequent exam, a student wrote: $\lim_{x \to 0+} \frac{5}{x} = \rotatebox{90}{5}$" (Frank 1990: A2). This is an over-reliance on syntactic manipulation, a misunderstanding of it. The lesson is that the rules of syntax must be spelled out; not *any* symbol manipulation goes.

What might the automorphism look like? There are two possibilities: It is the identity mapping, or it isn't. We will explore the latter case in §3.2.2. In the former case, the symbol manipulations (the syntax) that constitute the semantics are just those of the syntactic domain itself. This is the core meaning of understanding by "getting used to" the system (as in the syntactic way of understanding $\mathcal{L}'$ (§2.2.1)). One way to understand something is in terms of something else; another is to understand it in terms of itself. Both ways are important (if both are available).

Consider the way we learn algebra. We can learn it purely *syntactically* by learning rules for manipulating the symbols of algebraic equations: To find the value of x in '$2x + 4 = 6$', *move* the '4' from the left-hand side to the right-hand side, and *change* its *sign* (from '+' to '–', yielding '$2x = 6-4$'); then *move* the '2' from the left-hand side to the right-hand side, and *change* its *location* (from "above the line" to "below the line", yielding '$x = \frac{6-4}{2}$'); finally, *simplify* the right-hand side to yield the answer ('$x = \frac{6-4}{2} = \frac{2}{2} = 1$'). (Is "simplification" syntactic or semantic? Arguably the latter, but equally arguably the former: One memorizes the facts that $6 - 4 = 2$ and that $2/2 = 1$; table look-up is syntactic, or at least so I will take it here. Cf. Shapiro 1977.) The full set of these techniques permits one to solve any such equation; it was, in fact, the way I first learned algebra. We can also learn algebra *semantically*: To find the value of x in '$2x + 4 = 6$', *model* the equation as a scale with (a) two identical but unknown weights and four 1-unit weights in the left balancing pan and (b) six 1-unit weights in the right pan; always keeping the pans balanced, remove four 1-unit weights from each pan; then remove half of the remaining weights from each pan; the result is a balanced scale with one unknown weight in the left pan and one 1-unit weight in the right; hence, the unknown must weigh 1 unit. (Unfortunately, I was never taught this in school; I picked it up from watching "educational" TV. See my "Searle's Experiments with Thought" (Rapaport 1986c) for further discussion.) Full understanding comes from merging the syntactic and semantic modes of understanding: "Marvin L. Minsky ... likes to say that you do not understand anything until you understand it in more than one way" (Kay 1991: 147–148).

Is one of these ways "better" than the other? Minsky, no doubt, would say that both are needed (and that if there were a third way, so much the better). Stephen S. Willoughby, in *Contemporary Teaching of Secondary School Mathematics*, says that "in general it is better *to teach through understanding* [that is, semantically] the first time than *to teach by rote* [that is, syntactically] and then say 'Now would you like to know what you've been doing all of this time?' " (1967: 101; my italics and interpolations). Perhaps; though I successfully learned the rote syntactic way first, even though I did have an "Aha!" experience when I subsequently learned the semantic interpretation. In this situation, the "magic incantation" of the syntactic domain *sufficed*, since there was an isomorphism between the two domains. I didn't *need* to "understand" the syntactic domain, since I had reason to believe (or had faith?) that the syntactic moves "worked".

A different example might clarify this. Consider the "standard" (division-like) algorithm for computing square roots (cf. Willoughby 1967: 101–107, Levesque 1986: 84). One can learn it, but, even so, one hardly understands what it means or how it works. Even when one is told (or figures out) why it works (via a geometric analogy), so that one has the semantic domain with which to compare it, it's *still* hard to understand. (Moreover, the understanding of it is—arguably—almost entirely symbolic/syntactic: One comes to understand, for example, that certain numbers in the algorithm must be doubled because there is a '2' in the middle term of the expansion of $(a + b)^2$.) Yet one can compute square roots with it, and even know that they *are* square roots, without "understanding" it.

Moreover, if, in order to understand a syntactic domain, you must also understand the semantic domain, then you have *two* things to learn, not one, and, of course, this method only works if you antecedently understand the semantic domain. But perhaps the issue is really: What's being taught—the semantic domain or the syntactic domain? The "correct" answer is, probably, the *semantic* domain. After all, the syntactic domain is mere notation, a way of expressing the semantic domain in language (either the language of mathematics or "mathematical English"). Still, to understand the *semantic* domain, it's best to understand *it* by "getting used to it" (which is, I am arguing, a *syntactic* enterprise) and *then* to learn the standard syntactic way of expressing it.

The question of what is really being taught is the issue of what mathematics is: Is it syntax? That is, is it pure symbol manipulation, as the formalists tell us? "[M]athematics may be defined as the subject in which we never know what we are talking about, nor whether what we are saying is true" (Russell 1901: 71).[2] Or is it something else? I don't think I have to answer that in the present context, though perhaps I should record that I favor the formalistic answer: Mathematics is an "abstraction" whose semantic interpretation (the "something else") is an "implementation"—a model, in the model-theoretic sense. More on abstractions and implementations anon (Chapter 7).

James T. Cushing, in "Quantum Theory and Explanatory Discourse" (1991), seems to deny the sufficiency of "getting used to". He seems to argue that "psychological acclimation" (p. 346) is not sufficient to turn mere "explanation" in terms of deductive–nomological entailments (which is clearly syntactic) into "understanding". By 'understanding', he means a semantic "interpretation of the formalism that allows us to comprehend" it (p. 338; cf. p. 347) in terms of mental imagery

[2]Interestingly, one year later, he wrote that "Mathematics, rightly viewed, possesses not only truth, but supreme beauty ..." (Russell 1902: 57).

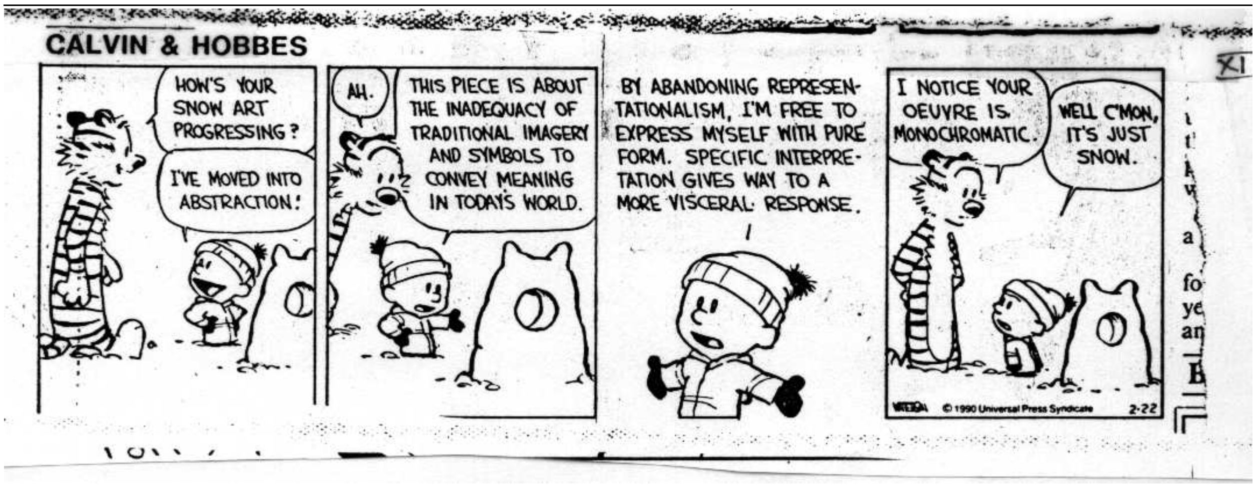


Figure 3.1:

or "picturability" (pp. 341, 343). He is willing to admit that it might just not be possible to "understand" certain scientific theories (for example, quantum mechanics)—that we can't get used to them. My contention is that he is too pessimistic, because he is too demanding. He is too demanding, because he requires understanding in terms of our *current* mental imagery. He does not seem to allow for the possibility that our imagery might change as we become more psychologically acclimated to a new syntax.

But even such understanding in terms of mental imagery is itself just a psychologically acclimated use of a syntactic system. Consider the work of Rohini Srihari on understanding captioned photos: The photo is understood by first understanding the caption, then constructing a "mental image" (as it were) of what the photo should be like, and finally matching that "mental image" against the actual photo, understanding the photo in terms of the "mental image". But when one looks at the details of how this is done, one sees that it is entirely syntactic. Thus, ultimately, it ought to be possible to understand a syntactic system in itself.

By the way, understanding by "getting used to"—syntactic understanding, or syntax for its own sake—has an interesting relationship to non-representational art, as in a *Calvin and Hobbes* cartoon (Figure 3.1). Representational art is supposed (or: intended) to represent something in the world, and in such a way that (almost) anyone[3] can recognize or understand it. Abstract art is pure syntax, pure form. (Thus, it has no content, a point to which we shall return (§3.2.2.2.3).)

[3]At least, those within the culture that the picture is part of. Pictures made by members of the Mparntwe Arrernte community in Alice Springs, Australia (and by other groups in Central Australia, such as the Warlpiri and the Pintupi), take a bird's-eye view and cannot be understood by non-Arrerntes without instruction. Moreover, Western paintings cannot be understood by *them* without instruction: A picture of a horse shown from the side is seen/interpreted by Arrerntes as a horse lying on its side, hence, presumably, dead! (David Wilkins, personal communication, 1990; cf. Wilkins 1991: 217.)

3.2.2 Using Parts to Understand the Rest.

If the automorphism of §3.2.1 is *not* the identity mapping, then it must map some elements onto others (or sets of others). So some parts of the syntactic domain will be understood in terms of others. (There may be "fixed points"—symbols that *are* mapped into themselves; we'll come back to those in §3.2.2.2.) Let's see what this means for our central case—natural-language understanding.

3.2.2.1 Dictionary definitions and algebra.

Dictionary-like definitions are an obvious example of this sort of automorphism. Indeed, this is probably what *most* people mean by "meaning", as opposed to philosophers, logicians, and cognitive scientists—though not some cognitive archeologists: "without going into a profound semiotic analysis, we can perhaps defining 'meaning' as 'the relationship between symbols' " (Renfrew 1990: 18). Now, this is bad semiotic analysis; semiotics tells us that meaning (semantics) concerns relations between symbols and the world—it is *syntax* that concerns relations among symbols. But Renfrew's view is fine on the theory being presented here: All correspondences—even purely syntactic ones—can be seen as semantic. Even some cognitive scientists are sympathetic: Yorick Wilks and Dan Fass claim that "meanings are, if anything, only other symbols" (Wilks & Fass 1992: 205; cf. Wilks 1971: 505, 506, 511); Wilks elaborates on this point:

> . . . except in those special cases when people do actually draw attention to the external world in connexion with a written or spoken statement, 'meaning' is always *other words*, and talk about 'the senses of words' is only a disguised restatement of that fact. (Wilks 1972: 86.)

And, as noted earlier (§2.8.2), we learn the meaning of many (if not most) new words in linguistic contexts—either in explicit definitions or "on the fly" in ordinary conversational or literary discourse. The unknown word, like the algebraic unknown, simply means whatever is necessary to give meaning to the entire context in which it appears. The meaning of the unknown word is (the meaning of) the surrounding context—the context "minus" the word. Finding the meaning is, thus, "solving" the context for the unknown. As James Higginbotham expresses it in "On Semantics", "The appearance of a word in a restricted number of settings suffices to determine its position in the language as a whole" (1985: 2; cf. Wilks 1971: 519–520). As Wilks notes in "Decidability and Natural Language" (1971), the context must be suitably large to get the "correct" or at least "intended" meaning. But, as the 'vase' example shows (§2.8.2), any context will do for starters. One's understanding of the meaning of the word will *change* as one comes across more contexts in which it is used (or: as the total context becomes larger); ultimately, one's understanding of the meaning will reach a stable state (at least temporarily—everything is subject to revision). Thus, learning a word is theory construction: One's understanding of the word's meaning is a *theory*, subject to revision. (For details and further references, see: Rapaport 1981; Ehrlich & Rapaport 1992, 1993, 1995; Ehrlich 1995.) To get a feel for how this might work, let's consider a few actual cases.

Case 1. Learning the meaning of a new word. The first time I read the word 'brachet' (in Malory's *Morte Darthur* (1470)), I did not know what it meant. (Do you?) Here is the context of that first occurrence:

> [T]here came a white hart running into the hall with a white brachet next to him, and thirty couples of black hounds came running after them with a great cry. (Malory 1470: 66.)

My first hypothesis (believe it or not) was that a brachet was a buckle on a harness worn by the hart. Note two things. First, this hypothesis goes beyond the very constrained algebraic picture I've been painting, but the algebraic metaphor is still a reasonable one: All that needs to be modified is the notion of context—here, I am extending the notion to include the background knowledge (including "world knowledge" and "commonsense knowledge") that I bring to bear on my understanding of the narrative (cf. Rapaport 1991a, Rapaport & Shapiro 1995). Second, it doesn't matter whether this hypothesis is good, bad, indifferent, or just plain silly. If I never see the word again, it won't matter, but, if I do, I will have ample opportunity to revise my beliefs about its meaning. Indeed, after 18 more occurrences of the term,[4] I stabilized on the following theory of the meaning of 'brachet': A brachet is a hound or hunting dog, perhaps a lead hound. Not bad, considering that the *Oxford English Dictionary* defines it as synonymous with 'brach', which means "A kind of hound which hunts by scent" (Simpson & Weiner 1989, Vol. 2, p. 1043).

Case 2. Revising the meaning of a word. This divides into two subclasses:

Case 2A. Revising the meaning of a word whose meaning is currently misunderstood. The case I have in mind is 'smite', which, at the time I read Malory, I had thought meant "kill". In many of the contexts in which this term occurred, my theory was not disconfirmed. Imagine my surprise, however, when I read this:

> ... Balan smote Balyn first; he put up his shield, smote Balan through the shield, and cut his helmet. Then Balyn smote him again with that unhappy sword, and well nigh felled his brother Balan. They fought there together till their breath failed. ... [A]ll the place where they were was blood-red. By that time they had each smitten the other seven great wounds; the least of them might have been the death of the mightiest giant in the world. Then they went to battle again, so marvelously that to hear of that battle was to doubt it, ... (Malory 1470: 59.)

I had two choices: Either believe that these passages from the *Morte Darthur* described magical events (which they did not), or revise my definition of 'smite' to something like "hit very hard". The latter worked.

Case 2B. Augmenting the meaning of a word that has multiple meanings. I have a well-entrenched belief that 'to dress' means "to put clothes on". Consider the following passage:

> Therewith two of them dressed their spears and Sir Ulfyus and Sir Brastias dressed their spears .... (Malory 1470: 15.)

Presumably, they did not put clothes on their spears. After ten such occurrences,[5] I was able to add a second definition to my theory of the meaning of 'dress': "To prepare (as a weapon for battle, or

[4]Fewer occurrences might have sufficed, since I did not revise my definition after *each* occurrence of the term, but only after groups of occurrences. It took only 10 such groups before the definition stabilized. (The protocols appear in Ehrlich 1995.)

[5]Eight "groups".

troops for battle)". (Further reflection on the clothing sense and on such terms as 'salad dressing' reveals that 'dress' has a very general, core meaning of "prepare (for use)".)

Let me stress that this is purely syntactic in the following sense: First, there is no external semantic domain: I did not see a brachet (or a picture of one), I did not see smiting going on, and I did not see what Sir Ulfyus et al. were doing when they were dressing their spears. Second, when I read the word (or when Cassie does), I build a mental representation of that word embedded in a mental representation of its context. These mental representations are part of the entire network of mental representations in my mind. Thus, the background knowledge I contribute is part and parcel of the mental representation of the new word in context. It is that system of mental representations that constitutes the syntactic domain in which is located "the meaning" of—that is, my understanding of—the word.

Representing meaning in such a dictionary-like network goes back at least to M. Ross Quillian's "Semantic Memory" (1967, 1968), though he was more concerned with merely representing the information in a dictionary, whereas I am concerned with representing meaning as part of a cognitive agent's entire complex network of beliefs. This is a brand of conceptual-role semantics, since I take the meaning of a word to be, algebraically, the role it plays in its context. It is also a holistic view of semantics. I'll come back to these issues in Chapter 4.

3.2.2.2 Understanding the parts

Another thing that using parts of the syntactic domain to understand the rest of it might mean is that those parts are primitives. Still, how are *they* understood? What do *they* mean?

First, how does this work? Smith pointed out that "a semantic domain may of course include its own syntactic domain" (1982: 10). And Cho's point (§2.7.1) that, in communication, the hearer might be identical with the speaker (when one "talks to oneself") suggests that I-qua-hearer am mapping my own concepts into themselves (the concepts of me-qua-speaker).

We saw, in §2.8.2, how the meaning of "linguistic" nodes can be given in terms of "non-linguistic" nodes. Note that this is a kind of referential meaning, except that the internal nodes "refer" to other internal nodes rather than to an independent and external "worldly" domain (cf. §2.8.1, above, and Wilks 1972, §2.32, esp. p. 87).

Wilks has proposed that the meaning of a word can be determined from a large enough surrounding linguistic context, especially if that context can include a dictionary to help resolve ambiguities (1971: 519). Besides Wilks's own implementation of such a methodology (cf. Wilks 1975), Wlodek Zadrozny and Karen Jensen (1991) describe such a system using an on-line dictionary: "Reasoning takes place in a three-level structure consisting of an ***object level***, a ***referential level*** and a ***metalevel***. ... The *referential level* ... consists of theories representing background knowledge .... [The] 'grounding' of logical predicates in other conceptual structures ..." (p. 177). Thus, words "refer" to other words, namely, those in the appropriate entry in the on-line dictionary.

3.2.2.2.1 Damasio. Antonio R. Damasio's theory of "time-locked multiregional retroactivation (1989ab) is a similar theory from a neuroscientific standpoint. It is worth exploring in some detail.

Damasio begins in good top-down fashion by considering "the experiences that are conjured up in recall and are used for recognition". These are things that are psychologically or functionally characterized; they are not necessarily *abstract*: they are *real* and *felt*. And he seeks "a neural architecture capable of supporting them" (1989a: 26). Contrast this methodology with (a) seeking a *computational* model of these experiences, (b) building a computer that behaves the same way, and (c) isolating some neural structure and *then* seeking to determine what behaviors it supports (a sort of reverse engineering). Method (a) would be the AI-analogue of Damasio's strategy, whereas (b) would not (necessarily) be of much interest, since mere input–output equivalence would hardly constitute an explanatory theory (cf. McCloskey 1991). Method (c) is risky, though sometimes the only practical alternative: Certainly, if we had a computer whose behavior we were ignorant of, (c) would be our only option, but in the case of the brain, (c) is useful only to the extent that its results match those of the top-down functional approach and to the extent that the isolated neural structures had "carved nature at the joints".

Damasio's theory, as it unfolds, will be seen to bear some resemblances to two other cognitive science theories: connectionism and Hector-Neri Castañeda's theory of guises.[6] It contrasts (or seems to contrast)—in some respects—with the sort of "local" representational theory of typical "classical" or "symbolic" systems such as SNePS/Cassie. Consider: "... perceptual experience depends on neural activity in *multiple* regions activated simultaneously, rather than in a *single* region where experiential integration would occur" (Damasio 1989a: 26, my italics). That sounds very connectionist (or, at least, distributed); it also sounds like philosophical bundle theories (such as guise theory), in which objects are taken to be "bundles" of properties, rather than single items.

We might wonder whether simultaneity is the *only* unifying feature. A hint of an answer—as well as additional guise-theory–like claims—may be found in these passages:

> The two critical structures in the proposed architecture are the fragment record of *feature-based* sensory or motor activity, and the convergence zone, an amodal record of the combinatorial arrangements that bind the fragment record as they occurred in experience. (Damasio 1989a: 26, my italics.)

We will see what 'amodal' means in a minute (cf. also Damasio 1989a: 46). 'Binding' is closely related to syntactic structure, as can be seen from the next quotation:

> There are convergence zones of different orders; for example, those that bind features into entities and those that bind entities into events or sets of events, but all register combinations of components or terms of coincidence or sequence in space and time. (Damasio 1989a: 26.)

Note, first, that the binding "mechanism" seems to play the same roles that the c-operator and/or consociation play in guise theory: Just as one (kind of) convergence zone "binds features into entities", the c-operator "binds" properties into guises, and just as another (kind of) convergence zone "binds entities into [mental representations of] events", consociation "binds" guises into larger structures that play some of the same roles in guise theory as events or propositions.

[6]For now, I will assume that the reader is familiar with both of these. Good surveys of connectionism are Graubard 1988; *Cognitive Science*, Vol. 9, No. 1 (1985); and—from a critical standpoint—Pinker & Mehler 1988; a useful tutorial is Knight 1989a, 1990. On guise theory, see, for example, Castañeda 1972, 1975, 1977, 1980, 1989a; cf. Rapaport 1991b.

Second, *mere* simultaneity seems not to be the only unifying feature: Spatial contiguity plays a role, too. Nonetheless, "there is no single site for the integration of sensory and motor processes. The experience of spatial integration is brought about by time-locked multiple occurrences" (Damasio 1989a: 27–28). In our terms, there are correspondences between things, but there is no unitary "joining" of them, and the correspondence is primarily one of co-temporality.

A couple of questions can be raised: (1) If two unrelated fragments "occur" simultaneously, are they experienced as a single object? The answer seems to be 'yes': "convergence zones can blend responses, that is, produce retroactivation of fragments that did not originally belong to the same experiential set .... When pathological combinations of input are reached, the zone malfunctions, for example, it may generate 'fantastic' ... responses" (Damasio 1989a: 47).

(2) Is the experience a semantic interpretation of the fragments? If so, then, in light of the answer to question (1), could this explain our ability to think of non-existent objects such as unicorns and round squares? Indeed: "The existence of abstract entities are criterion-governed conjunctions of features and dimensions present in ... concrete entities ..." (Damasio 1989a: 42). And: "Both the representation of abstract entities and of events are derived from the representation of concrete entities are are thus individualized on the basis of combinatorial arrangement ..." (p. 44).

The purely syntactic nature of Damasio's theory of meaning is evident in the following passage:

> In this proposal, and unlike traditional neurological models, there is no localizable single store for the meaning of a given entity within a cortical region. Rather, meaning is reached by widespread multiregional activation of fragmentary records pertinent to a stimulus, wherever such records may be stored within a large array of sensory and motor structures, according to a combinatorial arrangement specific to the entity. (Damasio 1989a: 28.)

That is, meaning is the result of "action"—manipulation of the symbols of the neural system.

Unlike SNePS/Cassie or other "classical" systems, however, (mental) representations seem not to be permanent records: "A display of the meaning of an entity does not exist in permanent fashion. It is recreated for each new instantiation" (p. 28). But this is puzzling: Does such a display exist at least temporarily? If so, where? And wouldn't such a location, albeit temporary, be a "localizable single store for the meaning of a given entity"? The notion of "recreation" has been suggested by others (for example, Clancey 1991). I find it hard to comprehend. To be *re*created, mustn't a pattern be stored somewhere, somehow?

What is crucial, however, and apparently unchallengeable, is the lack of any *single* location for representing an object:

> Current knowledge from neuroanatomy and neurophysiology of the primate nervous system indicates unequivocally that any entity or event that we normally perceive through multiple sensory modalities must engage geographically separate sensory modality structures of the central nervous system. Since virtually every conceivable perception of an entity or event also calls for a motor interaction on the part of the perceiver and must include the concomitant perception of the perceiver's somatic state, it is obvious that perception of external reality and the attempt to record it are a

> multiple-site neurophysiological affair. ... And the fragmentation that obtains from concrete entities is even more marked for abstract entities and events, considering that abstract entities correspond to criterion-governed conjunctions of dimensions and features present in concrete entities, and that events are an interplay of entities. (Damasio 1989a: 28–29.)
>
> The experience of reality, however, ... is not parcellated at all. The normal experience we have of entities and events is coherent and "in-register", both spatially and temporally. Features are bound in entities, and entities are bound in events. How the brain achieves such a remarkable integration starting with the fragments that it has to work with is a critical question. I call it the *binding problem* .... The brain must have devices capable of promoting the integration of fragmentary components of neural activity, in some sort of ensemble pattern that matches the structures of entities, events, and relationships thereof. (Damasio 1989a: 29.)

Thus, *apparent* unity in the world is perceived by means of fragments and internal complexity. We will see how, shortly. Note, though, that we have here a neuroscientific analogue of a Kantian epistemology: Our conceptual schemes allow us to make sense of—to categorize—noumena, something like what William James described as a "blooming, buzzing confusion" (James 1893: 488). (I remember once, at a basketball game, thinking that what I was "really" perceiving was a huge congeries of tiny isolated sensations of light, color, and noise, and that my mind (or brain) was somehow able to integrate these into a coherent experience of a basketball game.)

Damasio's "amodal" solution to the binding problem is to be contrasted with solutions in which "the components provided by different sensory portals are projected together in so-called *multimodal* cortices in which, presumably, a representation of integrated reality is achieved" (Damasio 1989a: 29, my italics). Suppose I perceive (to honor Wilfrid Sellars's favorite object) a pink ice cube. On a *non*-Damasio, "multimodal" theory, there would be some location in my brain and some representation in that location that corresponds to the actual object. For instance, my visual and tactile systems will signal that I have perceived something pink, something icy(-looking), something cubical, and a representation of a pink ice cube will be created at some *multi*modal site. In SNePS terms, this solution might represent the pink ice cube either by a structured individual node, such as `M4` in Figure 3.2, or, perhaps, by a base node about which assertions are made, as in `B1` of Figure 3.3. A Damasio-like, "amodal" representation might be like `M7` of Figure 3.4 or `M4` of Figure 3.5.

Damasio offers, in Kuhnian fashion, a new amodal paradigm due to lack of evidence for the multimodal picture (Damasio 1989a: 30–38, esp. pp. 35–36). On the new paradigm, the unity of experience is an illusion:

> An answer to this puzzle, namely the ability to generate an integrated experience in the absence of any means to bring the experience's components together in a single spatial meeting ground, might be a trick of timing. It would allow the perceiver or recaller to experience spatial integration and continuity in relation to sets of activity that are spatially discontinuous but do occur in the same time window, an illusory intuition. (Damasio 1989a: 38.)
>
> The integration of multiple aspects of reality, external as well as internal, in

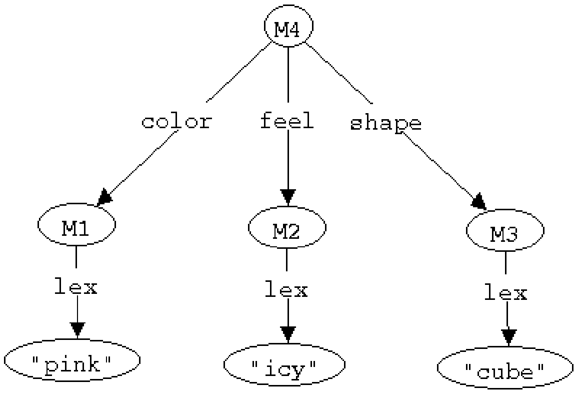


Figure 3.2: A non-Damasio, “multimodal”, SNePS representation of a pink ice cube as a structured individual; `M4` = a thing that is pink, icy, and a cube.

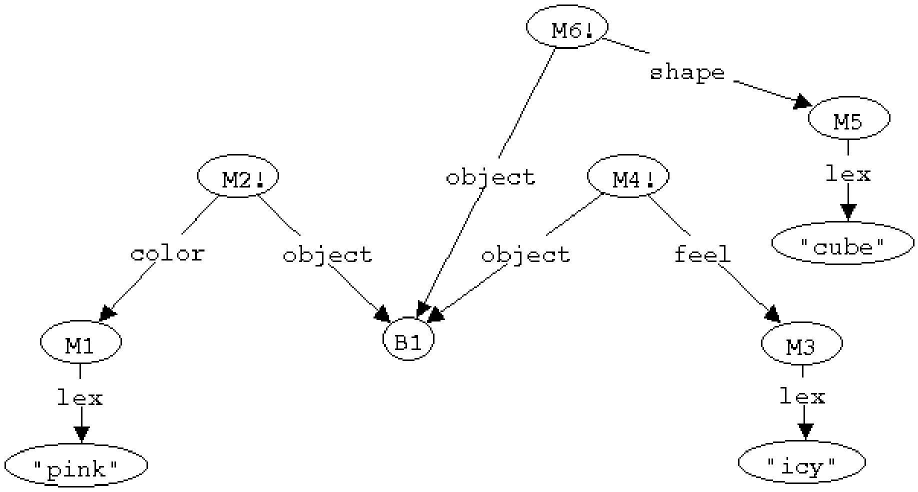


Figure 3.3: A non-Damasio, “multimodal”, SNePS representation of a pink ice cube as a base node (`B1`) about which three things are asserted: `M2!` = `B1` is pink; `M4!` = `B1` is icy; and `M6!` = `B1` is a cube.

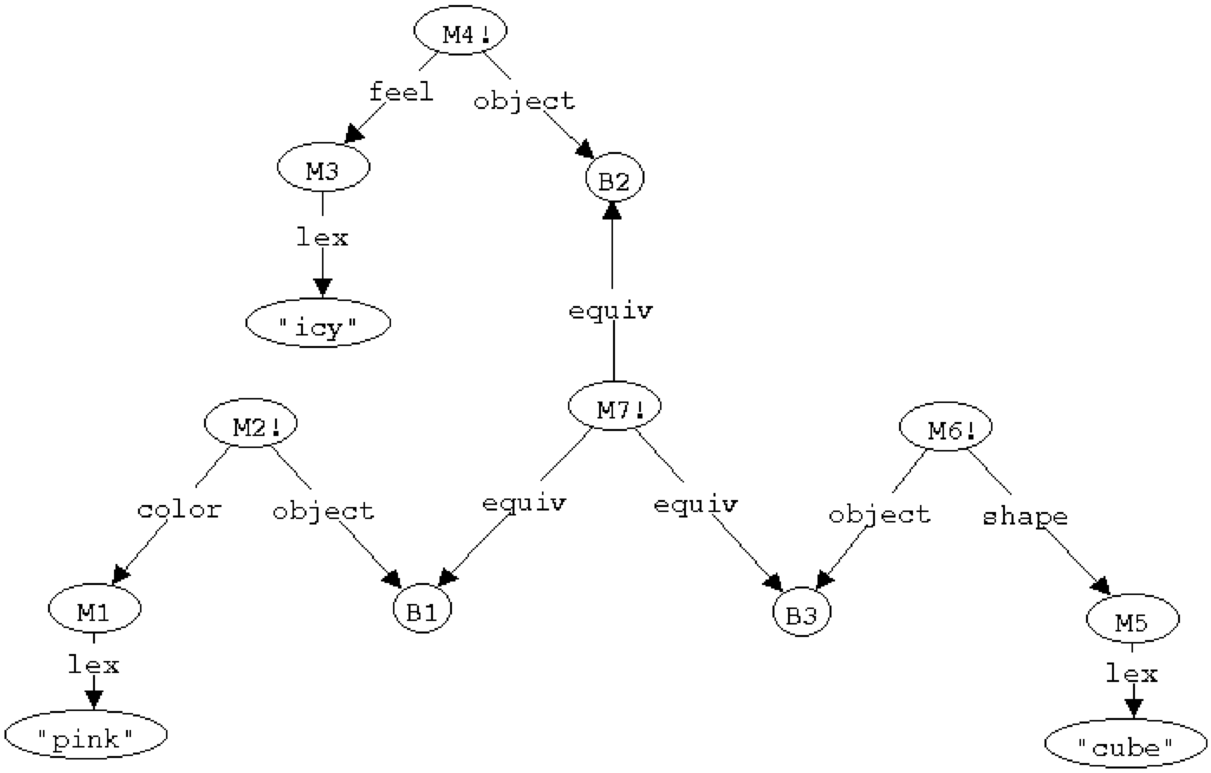


Figure 3.4: A Damasio-like, “amodal”, SNePS representation of a(n extensional) pink ice cube as three (intensional) individuals that are co-extensional:

`M2!` = `B1` is pink;
`M4!` = `B2` is icy;
`M6!` = `B3` is a cube;
`M7!` = `B1`, `B2`, and `B3` are “equivalent” (i.e., are the same (extensional) object).

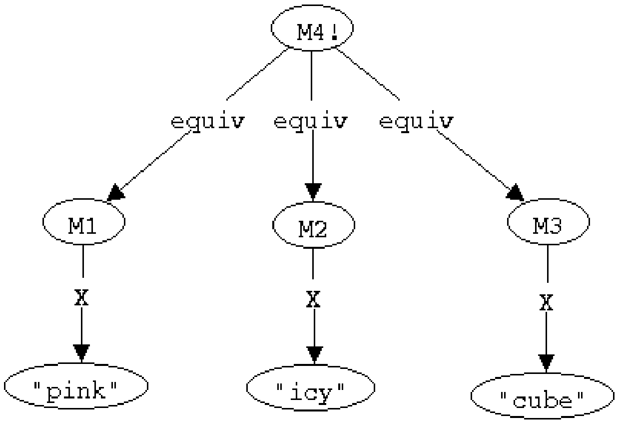


Figure 3.5: Another Damasio-like, "amodal", SNePS representation of a pink ice cube. `M4` represents something like the following: There is a temporally simultaneous experience of pinkness, iciness, and cubicness. (Here, the `X`-arcs are neither `lex`- nor `pic`-arcs, but some sort of generic sensory arc.)

> perceptual or recalled experiences, both within each modality and across modalities, depends on the *time-locked co-activation* of geographically separate sites of neural activity .... (Damasio 1989a: 39, my italics.)

So, co-temporality does appear to suffice for unifying or integrating the bundles of features (and is, hence, a plausible interpretation or implementation of Castañeda's c-operator). However, in another paper, Damasio says that "co-attention" is also necessary (1989b: 24–25).

What is curious, though, is that the features are located in one place, while their syntactic structure is recorded elsewhere:

> The representations of physical structure components of entities are recorded in precisely the same neural ensembles in which corresponding activity occurred during perception, but the combinatorial arrangements (binding codes) which describe their pertinent linkages in entities and events (their spatial and temporal coincidences) are stored in separate neural ensembles called *convergence zones*. (Damasio 1989a: 39.)

How are the combinatorial arrangements determined? How are they abstracted away from their content (so to speak)? And how are they linked together? Perhaps by what Damasio calls "reciprocal projection" (Damasio 1989a: 39): "The concerted reaction of physical structure fragments, on which recall of experiences depends, requires the firing of convergence zones *and* the concomitant firing of the feedback projections arising from them" (Damasio 1989a: 39, my italics). Thus, the neural ensembles recording the physical structure components (the fragments) send signals to the neural ensembles (the convergence zones) whose informational content is the combinatorial structure of the fragments, and the convergence zones send signals to the other neural ensembles, whose informational content allows the fragments to be "reconstructed".

The abstraction question I raised a moment ago has its answer in another curious feature:

> Convergence zones bind neural activity patterns corresponding to topographically organized fragment descriptions of physical structure, which were pertinently associated in previous experience on the basis of similarity, spatial placement, temporal sequence, temporal coincidence, or any combination of the above. (Damasio 1989a: 39.)

So the answer to the binding problem is twofold: In original perception, it's just the associations (by similarity, etc.); it is, therefore, an illusion. In recall, it's that plus the convergence zones.

There is a small difference between Damasio's theory and Castañeda's: For Damasio, the structure of (a mental representation of) an entity or event is not so much $c\{F_1, \ldots, F_n\}$, as Castañeda would have it, but: $\{F_1, \ldots, F_n, c\}$—that is, it's not that the convergence zones *operate* on the fragments, but that the fragments and the convergence zone are activated together. In Damasio's words, "The co-occurrence of activities at multiple sites, which is necessary for temporary conjunctions, is achieved by iteration across time phases" (Damasio 1989a: 40).

The other theory that Damasio's theory brings to mind is Fodor's language of thought:

> 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 activation 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.)
>
> 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 the binding code and it is based on a device I call the convergence zone. (Damasio 1989a: 45.)

So fragments correspondence to lexical items, and the binding code corresponds to syntactic structure. This is also very compositional: The representation of an object consists of features plus a combinatorial arrangement. Moreover, some representations are "linguistic" and some "non-linguistic" in what appears to be precisely the sense I described above and discussed in conjunction with Zadrozny and Jensen:

> The brain not only inscribes language constituents but also provides direct and dynamic neural links between verbal *representations* and the representations of non-language entities or events that are signified by language. In other words, the brain embodies (materializes)[[7]] in neural hardware the combined biological and cultural bond that culture has assigned between a language representation (a signifier) and a segment of non-verbal reality (a signified) .... (Damasio 1989a: 55; my italics.)

[7] I would say 'implements'—WJR.

3.2.2.2.2 The symbol-grounding problem. We are still left with the problem of how those parts in terms of which the rest of the network is understood (for example, the "non-linguistic nodes") are themselves understood. Three possibilities suggest themselves.

First, our automorphism might have no fixed points. This case leads quickly to a holistic theory in which the meaning of a node is—ultimately—definable by its location in the entire network. We'll discuss this in Chapter 4.

Second, if the automorphism *has* fixed points, they might be "markers" that have no intrinsic meaning. But such markers *get* meaning the more they are used—the more roles they play in providing meaning to *other* nodes. A helpful analogy comes from Wartofsky:

> But I would argue that ... 'mental' objects, or 'internal representations' are derivative, and have their genesis in our primary activity of representing, in which we take external things,—most typically, what we also designate as physical objects—as representations. Moreover, I take our *making* of representations to be, in the first place, the actual praxis of creating concrete objects-in-the-world, *as* representations; or of taking the made objects as representational. (1979: xxi–xxii.)

Of course, this is a *claim*, not an argument. But in this *primary* activity, what do we take the external physical object to be a representation *of*? Its use and history? What does that have to do with the common properties in terms of which one thing can represent another? More likely, it is that, once made, it can *remind us* of its use or of its manufacture and therefore represent those things for us. The *first* time we see an unfamiliar object, it is a meaningless thing (except insofar as it shares any properties with anything familiar, allowing us to form hypotheses about it and to place it in our semantic network). The *second* time we see it, it can remind us of something, if only of itself on its first appearance. Subsequent encounters produce familiarity, which entrench it in our network, and allow newer objects to be understood in terms of *it*. This is how holism works. Thus, our first two possibilities lead to a holistic conceptual role semantics; to repeat, we'll discuss that in detail in Chapter 4.

The third possibility is that the fixed points (or the markers, or—for that matter—any of the nodes) are somehow "grounded" in another domain. This, of course, is just to say that they have meaning in the correspondence sense of semantics, and ultimately we will be led to question the way in which we understand that other domain. But let's look briefly at this problem, which Stevan Harnad has dubbed the "symbol-grounding problem".

3.2.2.2.2.1 *The circular dictionary.* The symbol-grounding problem, in a nutshell, is the problem we have just been dealing with—how a hermetically sealed circle of nodes can be "grounded". The problem, according to Harnad, is that without such grounding, there can only be circular meaning. And, presumably, circles are vicious and to be avoided.

Consider a dictionary. It is well known that it is a closed circle of meanings. Indeed, it should be obvious: After all, each word is defined in terms of other words. Assuming that all words used in the definitions are themselves defined, we have a circle (in fact, several of them). Now, before agreeing with Harnad that such circles cannot yield meaning or understanding, consider that we do use dictionaries fairly successfully to learn meanings. How can this be? If the circle is a very small one (say, 'being' is defined as "existence", and 'exist' as "have being"),[8] then it may indeed not be informative, especially if we don't antecedently know the meaning of the definiens. However, the larger the circumference of the circle, so to speak, the more likely it is that it will be informative, on the assumption that the definition of the word whose meaning we seek, and the definitions of the words in that definition, and so on, will contain *lots* of words that we antecedently understand. So, we can easily "solve" the "equation" for the unknown word—that is, dictionary definitions are most useful to the extent that they are like '$x = \frac{4-3}{2}$', rather than like '$x = y$' or '$x = \frac{4y-3}{2}$' (that is, where there is a further unknown in the definiens).

Nonetheless, some words will still only be ill-defined in terms of other words, notably (but not exclusively) nouns like 'cat' or 'cow'. For these, *seeing* a cat or cow (or for 'love', *experiencing* love) is worth a thousand-word definition. Granted—the meaning of 'cat' is better grounded in a perceptual experience of a cat than in other words. Now, some dictionaries go a bit further: They are illustrated. This helps considerably. We have then a type-distinction in the syntax of the definiens: Terms can be of the type *word* or the type *picture*. Note, however, that, although we now have grounding in an extra-linguistic system, *it is still part of the dictionary*. And, of course, the pictures could, with a suitable indexing scheme, themselves be definiendum entries: A picture of a cat could have as its "definiens" the word 'cat', as in a visual dictionary or a field guide to cats. This only widens the circle, as I'm sure Harnad would be quick to point out. We could widen it further, albeit at some expense and inconvenience: Let every dictionary come with a real cat; ditto for all other better-ostensively-defined terms. We still have a circle, but now, I think, Harnad would have to agree that we've also got grounding—we've merely incorporated the groundings into the dictionary.

The same holds for the mind. What Harnad says is needed is a link between (some) mental nodes (say, our "cat" node) and items in the external world (say, a cat). Now, it's true that we can't import such items directly into our minds. What we *can* do is have mental representations of those items. And it is the relation between our "cat" node and a node representing a perceived cat that grounds the former. We saw how in §2.8.2: What we (or Harnad) *think* is the relation between word and world is really a connection between an internal representation of a perceived *word* and an internal representation of the perceived *world*.

There is no doubt in my mind whatsoever that experience enriches our understanding. Consider "immersion" learning of a foreign language, say, French. "Thinking in French" is understanding French holistically, without any correspondences to one's native language (say,

[8]According to *Webster's Vest Pocket Dictionary* (Springfield, MA: Merriam-Webster, 1989), 'being' means "existence", and 'exist' means "have real or actual being", so 'exist' means "have real or actual existence". Worse, 'real' means "actually existing", so 'exist' means "have actually existing, or actual, existence". Worse yet, 'actual' means "really existing", so 'exist' means "have really existingly existing, or really existing, existence".

An even shorter circle can be found: 'proof' means "evidence of a truth or fact"; 'evidence' means "proof or testimony", and 'testimony' means "statement given as evidence".

Shorter still: 'realize' means "be aware"; 'aware' means "having realization or consciousness"; 'conscious' means "aware".

For what it's worth, the *Oxford English Dictionary*, *Webster's New International Dictionary* (second edition), *The American Heritage Dictionary* (second college edition), and *The Random House Webster's College Dictionary* all define the number words 'two' through 'ten' roughly recursively (as meaning the preceding number + one) and they define 'one' as the first cardinal number (though the *OED* also points out that one is the number that when added to itself yields two). *Webster's 9th New Collegiate Dictionary* has as its first definition of, for example, 'ten', an ostensive one (a reference to a not-very-helpful number table) and as its second definition, "the tenth in a set or series", which, arguably, is a *very* small circle indeed!

English). It is helped immeasurably by living in a francophone community. When we ask "What does (the French word) '*chat*' mean?", and we give the answer ("cat") in English words, we are doing pure *syntax* (here, relating symbols from one system to those of another) *that is also semantic* (understanding one system in terms of another). This is no different than answering the question in French ("*un chat est un petit animal domestique, dont il existe aussi plusieurs espèces sauvages*")[9]—except for choice of language for the definiens (and a certain verbosity necessitated by staying within the French "circle"). Giving the definition in English is just as much symbol grounding as pointing to a cat would be. Symbol grounding, thus, does *not* necessarily get us out of the circle of words—at best, it widens the circle. And that is the point I want to make: Syntactic understanding—the base case of understanding—is just a *very* wide circle.

3.2.2.2.2.2 ***Harnad's theory of symbol grounding.*** In fact, Harnad's own examples of grounding are internal in just the ways we have been considering:

> A candidate solution ...: Symbolic representations must be grounded bottom-up in non-symbolic representations of two kinds: (1) *iconic representations*, which are **analogs** of the proximal sensory projections of distal objects and events, and (2) *categorial representations*, which are **learned and innate** feature detectors that pick out the invariant features of object and event categories from their sensory projections. Elementary symbols are the names of these object and event categories, assigned on the basis of their (nonsymbolic) categorial representations. Higher-order (3) *symbolic representations*, grounded in these elementary symbols, consist of symbol strings describing category membership relations .... (Harnad 1990: 335, Abstract; Harnad's italics, my boldface.)

Harnad distinguishes between "symbolic" and "non-symbolic" representations. But *both* are *internal* representations! Harnad says that the non-symbolic "*iconic representations* ... are **internal** analog transforms of the projections of distal objects on our sensory surfaces" (p. 342; Harnad's italics, my boldface). A "projection of [a] distal object on our sensory surface" could be a retinal image, say. So an iconic representation is some "analog transform" of *that*, stored (or created) somewhere else (further along the optic pathway). This internal representation can be part of our semantic network (cf. Srihari 1991b).

Furthermore, the two representational systems (symbolic and non-symbolic—or perhaps there are three: symbolic, iconic, and categorial) must be linked; hence, because of Smith's Gap, they must all be internal. Clearly, the distinction between the two systems *can* be made, but to what end? Elementary symbols might correspond to nodes at the tails of `lex` or `pic` arcs (cf. §2.8.2); iconic and/or categorial representations might correspond to nodes at the heads of such arcs. They are all part of a single, albeit typed, representational system. That is, they are all terms in a formal syntactic system. (For details of how such a typed system works, one with both purely symbolic (or "linguistic") and "non-symbolic" (or "pictorial"—perhaps even "iconic" in Harnad's sense), cf. Srihari 1991b.)

Where is the "grounding"? One would *expect* internal items to be "grounded" in *external* ones. But look at Harnad's hierarchy of items (p. 335, Abstract): Symbolic representations are

[9] *Dictionnaire de Français* (Paris: Larousse, 1989): 187. Translation: A cat is a small domestic animal of which there also exist many wild species. Hardly an adequate definition!

"grounded" in "elementary symbols", which are "names" of "categories", which categories are "assigned on the basis of" categorial representations, which representations are "derived" from sensory projections; and iconic representations are "analogs" of those sensory projections. (This can't be strictly correct: surely, categorial representations must be derived from iconic ones. And the categories *themselves*—the ones named by elementary symbols and assigned on the basis of categorial representations—seem otiose: The symbols could name the categorial representations.) Note that the only actual use of the term 'grounding' is between symbolic representations and elementary symbols, *both of which are internal*. Indeed, *all* the items on this hierarchy are internal!

Moreover, what Harnad calls "non-symbolic" representations are, on his own terms, symbolic (or else his definition of 'symbol system' is too stringent, ruling out things, such as SNePS, that clearly are symbol systems). According to Harnad, for something to be a symbol system, it must *inter alia* be "(1) a set of arbitrary *physical tokens* that are (2) manipulated on the basis of *explicit rules* ..." (Harnad 1990: 336). So, arbitrary tokens that are *not* manipulated by rules are not symbols. But surely the iconic and categorial representations are, or can be, manipulated (as in Srihari 1991b). The rules must also be (strings of) physical tokens (p. 336). Surely, though, the rules could be *implicit*, that is, not part of the representational system. For instance, the inference rules that manipulate logically the symbols in a SNePS network are part of the SNePS Inference Package, not explicitly represented in the SNePS network itself. Similarly, as Lewis Carroll noted (1895), the rules of inference of a natural deduction system are not among the wffs of the system (though, to be fair, Harnad seems willing to accept this; cf. Harnad 1990: 336fn1).

Further, "The ... manipulation is based (4) purely on the *shape* of the symbol tokens (not their 'meaning'), i.e. it is purely *syntactic* ..." (p. 336). But this could hold of the "non-symbolic" representations unless they had *no* shape; yet Harnad says that they have "nonarbitrary shapes" (Harnad 1990: 335, Abstract). As long as they have shapes, they can be syntactically manipulated. Finally, "the system can be *systematically* assigned a meaning (e.g. as standing for objects, as describing states of affairs)" (Harnad 1990: 336). Now, this has to be an external or purely referential semantic interpretation. In any case, it holds for iconic representations, too: Surely they stand for objects. It is not so clear what categorial representations would stand for: If Lakoff (1987) is right, there *aren't* any categories out there. If categorial representations, then, can't be given a model-theoretic semantic interpretation, then they would in fact be *purely* syntactic. (They would also be intensional, in the sense that what they are "about" don't exist.)

Let's look at the category problem more closely:

> So we need horse icons to discriminate horses, but what about identifying them? Discrimination is independent of identification. I could be discriminating things without knowing what they were. ... For identification, icons must be selectively reduced to those *invariant features* of the sensory projection that will reliably distinguish a member of a category from any nonmembers with which it could be confused. Let us call the output of this category-specific feature detector the *categorical representation*. (Harnad 1990: 342.)

Of course, if Lakoff is right, this will be much more complex, and there might not be such invariants (but rather a family of them, etc.). Nonetheless, we humans can and do discriminate in more or less this way, building categories, which might be Lakoffian "idealized cognitive models". But idealized cognitive models are complex, symbolic (maybe), and highly interconnected. (The 'maybe' hedge on 'symbolic' has to do with Lakoff's apparent claim that his theory is not computational; cf. Lakoff 1987, e.g., pp. 343–345.) However, although it *might* not be "classically" computational, it yet might be connectionistically computational, which is all that is needed for my purposes. (There is no reason to think that it could *not* be "classically" computational; but arguing for that would lead us astray.) So there is no reason, not even on Harnad's own terms, to think that his system of "symbolic" and "non-symbolic" representations is not *entirely* symbolic.

Now, Harnad says that the "iconic and categorical representations are nonsymbolic" because "[t]he former are analog" and "the latter are icons" (p. 342), hence also, presumably, analog. But why does this make them non-*symbolic*? They are physical tokens, and surely, as such, they can be manipulated syntactically on the basis of rules. (By the way, Lakoff would disagree that categorical representations would be icons, since that implies that there are categories *in the world* for them to be icons of. If they're *not* icons, then they *could* be symbolic.)

Curiously, Harnad only mentions "grounding in the world" in a footnote:

> If a candidate model [for a cognitive system] were to exhibit all … behavioral capacities, both *linguistic* ["produce" and "respond to descriptions of … objects, events, and states of affairs"] … and *robotic* ["discriminate, … manipulate, … [and] identify … the objects, events and states of affairs **in the world they live in**] …, it would pass the "total Turing test" …. … A model that could pass the total Turing test, however, would be **grounded in the world**. (Harnad 1990: 341fn13; Harnad's italics, my boldface.)

Recall the "blind" blocks-world robot and the Rochester checkers-playing robot (§2.7.1). The former is blind and methodologically solipsistic. The latter can see. But is it grounded? Could it be fooled as the blind robot was? Possibly: by deceiving its eyes (shades of Descartes!). (And shades of *Star Trek*'s "Menagerie" (or "Cage") episode, in which aliens deceive Captain Pike into believing all sorts of things that are not real.) It would then occupy a world in which to be was to be perceived. Of course, such a Berkeleyan robot *would* be grounded *in the world that it lives in*, which happens not to be the actual world, but a purely intentional one. (In this case, note that the grounding system *and* the grounded one are *both* internal, hence part of a single network.) What would such a robot's symbols mean *to it*? Here, internal semantic interpretation would be done by internal links only.

We could go a step further and imagine that this robot is in fact behaving in the real world, only it's not playing checkers (perhaps it's discussing war strategies or lattices, or proving theorems). *We* would interpret what it's doing very differently from what *it* would. But if there are no "disagreements", then how would we or it know the difference?

Even more curious is the fact that grounding for Harnad—even grounding in the external world—does not seem to serve a semantic function:

> Iconic representations no more "mean" the objects of which they are the projections than the image in a camera does. Both icons and camera images can of course be *interpreted* as meaning or standing for something, but the interpretation would clearly be derivative rather than intrinsic. (p. 343.)

Harnad seems to be saying here that the causal connection of the iconic representations with its real-world counterpart is irrelevant to its intrinsic meaning. In that case, Harnad owes us answers to two questions: (1) what does such a causal grounding *do* in his theory, and (2) what *is* the intrinsic meaning of an iconic representation? Harnad may have identified an interesting problem, but he doesn't seem to have solved it.

My position is this: The mind–world gap cannot be bridged by the mind. There are causal links between them, but the only role these links play in semantics is this: The mind's internal representations of external objects (which internal representations are caused by the external objects) *can* serve as "referents" of other internal symbols, but, since they are *all* internal, meaning is in the head and is syntactic.

3.2.2.2.3 The body as ground. Let's suppose, however, for the sake of the argument, that external grounding *is* to be sought. Where shall we seek it? In unconstrained reality? Perhaps; though why wouldn't that lead to unconstrained misunderstanding? After all, if each of us grounds our concepts in arbitrary parts of the external world, what is to guarantee that there is any overlap in our several groundings? Better to ground our concepts in something common: our bodies.

Let me make the same point in a slightly different way. We have been considering the question "What does it mean to understand a system in terms of itself?", and the answer I have been favoring is that we understand such a system syntactically. But this answer seems to lead either to holism (which may not be a bad thing, though some think it is, because of its alleged circularity) or to taking certain symbols in the system as primitive (hence *not* understandable, except perhaps retroactively in terms of their contribution to the meanings of the *other* symbols, which leads us straight back to holism; cf. Hill 1994, 1995).

So maybe there *is* some "distinguished" or "marked" domain that has the following features:

1. Other domains can be understood in terms of it, recursively, so that *it* is the fundamental domain of understanding, so to speak.

2. It wears its own semantics on its sleeve, so to speak. That is, it is neither the case that it must be understood in terms of something else nor that it must be understood syntactically. Rather, it has some kind of intrinsic or original semanticity.

Now, I don't understand what feature (2) could possibly mean other than what I've been calling 'syntactic understanding', and maybe that's all it is. But what I've been calling 'syntactic understanding' is capable of being had by *any* domain, whereas the domain that would satisfy (1) and (2) is supposed to be special in some way. Can we, at least for the sake of argument, identify such a domain?

Yes—the body. Or, to be more specific, the human body (and, to be even more specific, in my case it would be *my* body). So the general idea is that at least some (if not all) of the concepts of my language (or my language of thought) are to be understood in terms of (that is, correspond to) parts or features of my body. So, almost all understanding is of the first kind—semantic or model-theoretic understanding, with my body as the ultimate or foundational semantic domain.

Fine. How, then, do we understand our bodies? My answer: We get used to them! Actually, this has to be *everyone's* answer. What I claim is that if there's *one* domain that is (in fact, can only be) understood by getting used to it, then *any* domain can be so understood. And, since I'm not solely interested in how *humans* understand, but rather in how *any* cognitive agent, including computers, can understand, I'm not especially interested in one special case. Nonetheless, it's certainly plausible that our bodies play this important and perhaps unique role in *our* case and, moreover, that an arbitrary cognitive agent's body plays the analogous role for *it*—no matter what that body looks like. I'll have more to say about the general case later (§3.2.2.2.4). (For a philosophical science-fiction investigation of this, see Justin Leiber's *Beyond Rejection* (1980).)

One nice thing about our bodies: They're always with us! They are a nice, handy [sic!], portable standard for grounding other concepts. Moreover, we have mental representations (images) of the parts of our bodies, and we have visceral feelings of how to manipulate the parts. Recall the *Calvin and Hobbes* cartoon (Fig. 3.1): If there is no "intended interpretation" of some domain (as in the case, say, of non-representational art), we will try to interpret it in any way we can—perhaps in terms of our bodies as a sort of default case. Zadrozny and Jensen cite Michael Turner's *Death Is the Mother of Beauty*, in which he suggests that the meanings (the semantics) of some terms are "constrained by our *models of ourselves* and our worlds" (Turner 1987: 7, cited in Zadrozny & Jensen 1991: 177; my italics). Our internal self-model, which must include a model of our body, can be directly understood in terms of our body *itself*. But our body, we have just seen, is a domain consisting of items related in certain ways and manipulable in certain ways. So it is a syntactic domain.

That's the general idea. Specific versions of it differ in detail or emphasis. Let me just cite a few, without going into their details:

1. There is, first and perhaps foremost, George Lakoff's and Mark Johnson's theory in which the body is the source of most metaphors and of the fundamental idealized cognitive models that structure our language and thought (Lakoff 1987, Johnson 1987). As an example, consider my use of 'handy' in the previous paragraph.

2. There are the very general and interrelated phenomena of indexicality and "situatedness". For instance, David Kirsh notes that "systems often think about the world indexically, in an *egocentric* fashion, which cannot be adequately interpreted in terms of properties of objective space time regions" (Kirsh 1991: 21). Such thinking is clearly body-centered (contrast Nils J. Nilsson's discussion of context-*free* knowledge (1991: 33)).

3. Another case, which shares features with both of the others (and may be subsumable in terms of them), is that of Helen Keller after the well-house episode: "As we returned to the house every object which I touched seemed to quiver with life. That was because I saw everything with the strange, new sight that had come to me" (Keller 1905: 36). She had a knowledge of her surroundings that she antecedently understood in relation to herself and her body (for example, tactally), which grounded her newly understood language.

4. Aaron Sloman raises the "semantic linkage problem", the problem of how a cognitive agent who uses a symbol S to refer to an object O can "relate" to O other than by S. He concludes "that when O is *part* of [the agent] ..., the link may be a comparatively simple causal relationship" (1985: 996). Although Sloman does not explicitly cite the agent's body as a "part", clearly it could be.

Now, one sort of objection that can be raised against this sort of view is akin to one raised against the mind–brain identity theory. If mental states and processes are identified with *human* brain states and processes, then—by definition—only humans can think. This seems rather chauvinistic. Functionalism gets around this by allowing for the possibility that mental states and processes are to be correlated (if not identified) with physical states and processes, leaving open what the possible physical media are. (Or even nonphysical, if one wants to include, say, angels; cf. Fodor 1981: 114.) Similarly, the lessons of Lakoff, of Johnson, and of the importance of indexicality need not be lost if we generalize beyond the *human* body. When Hubert Dreyfus, for example, argues that computers will never be able to think because they don't have bodies (1992, Ch. 7)[10] or aren't part of human society ("Introduction to the Revised Edition"),[11] he is overly pessimistic. He *may* have a point, but if he does, his point is that in order for a cognitive agent to think, it must have *a* body (to serve as foundational semantic domain) and be part of *a* society of, presumably, like-bodied (and like-minded) cognitive agents.

This suggests an interesting research project in robotics. The research project has two parts. Part 1 is to develop a computational cognitive agent that thinks in terms of the *human* body—eventually, it should be implemented in a human-body–like device (that is, it should be an android, if only on the order of *Star Wars*'s C3PO). As Nicolas Goodman has suggested (personal communication), "A computer that could understand human language would have to lead a pretty good simulacrum of a human life". Part 2 is to develop a computational cognitive agent implemented in a *non*-human body—that is, a robot, perhaps along the lines of *Star Wars*'s R2D2—and have it think in terms of *its* body.

3.2.2.2.4 Winston's problem. A need for a body also raises a serious problem: Cognitive agents with different (types of) bodies would have different concepts (we would literally be thinking different things, as in Figures 3.6 and 3.7). But these concepts would be thoughts nonetheless, and such differences might make mutual comprehension impossible. I will call this 'Winston's Problem', in honor of an early formulation of it by Patrick Henry Winston:

> Simulation of human intelligence is not a primary goal of this work. Yet for the most part I have designed programs that see the world in terms conforming to human usage and taste. These programs produce descriptions that use notions such as left-of, on-top-of, behind, big, and part-of.
>
> There are several reasons for this. One is that if a machine is to learn from a human teacher, then it is reasonable that the machine should understand and use the same relations that the human does. Otherwise there would be the sort of difference in point of view that prevents inexperienced adult teachers from interacting smoothly with small children.

[10] "If the body turns out to be indispensable for intelligent behavior, then we shall have to ask whether the body can be simulated on a heuristically programmed digital computer. If not, then the project of artificial intelligence is doomed from the start" (Dreyfus 1992: 235); and, of course, Dreyfus argues that the body *is* thus indispensable and *not* thus simulable.

[11] "... it [Winograd's SHRDLU] still wouldn't understand, unless it also understood that it (SHRDLU) couldn't own anything, since it isn't a part of the community in which owning makes sense. Given our cultural practices which constitute owning, a computer cannot own something any more than a table can" (Dreyfus 1992: 13); for my commentary on this, see Rapaport 1988, §4.2.

Interestingly, Dreyfus cites Herbert Simon in support of his point. Although Simon does say that SHRDLU "doesn't understand what it is to own something" (Simon 1977: 1061), he goes on to agree with the point I am making: "SHRDLU would understand what it meant to own a box if it ... could perform those tests and actions that are generally associated with the determination and exercise of ownership in our law and culture" (Simon 1977: 1061).

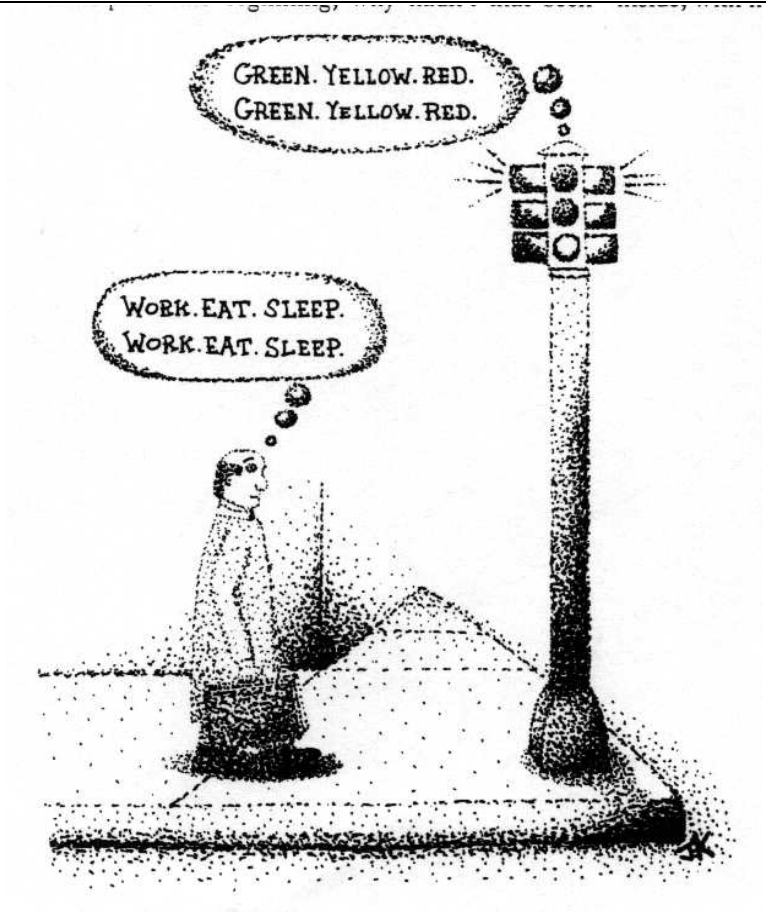


Figure 3.6: A *New Yorker* cartoon illustrating Winston's Problem.

Figure 3.7: "How birds see the world." (A *Far Side* cartoon illustrating Winston's Problem.)

> Moreover, if the machine is to understand its environment for any reason, then understanding it in the same terms humans do helps us to understand and improve the machine's operation. Little is known about how human intelligence works, but it would be foolish to ignore conjectures about human methods and abilities if those things can help machines. Much has already been learned from programs that use what seem like human methods. There are already programs that prove mathematical theorems, play good chess, work analogy problems, understand restricted forms of English, and more. Yet, in contrast, little knowledge about intelligence has come from perceptron work and other approaches to intelligence that do not exploit the planning and hierarchical organization that seems characteristic of human thought.
>
> Another reason for designing programs that describe scenes in human terms is that human judgment then serves as a standard. There will be no contentment with machines that only do as well as humans. But until machines become better than humans at seeing, doing as well is a reasonable goal, and comparing the performance of the machine with that of the human is a convenient way to measure success. (Winston 1975/1985: 143; cf. Kirsh 1991: 22–24).

(A version of Winston's Problem arises in those connectionist models in which it is not at all clear what, if anything, the final weights on the connections "mean" in terms of the task that the system has learned.) There are reasons to be optimistic, however. For one thing, Winston's Problem occurs, on a smaller scale, close at home: I, as a male, can never experience pregnancy; so, my understanding of 'pregnant' is qualitatively different from that of a female (certainly from that of a female who has been pregnant). Yet I use the word, am not misunderstood when I use it, and can understand (within recognized limits) a woman's use of it. (The example is Shapiro's; cf. Rapaport 1988: 116, 126n20.) For another thing, as long as two cognitive agents are communicating about a common (external) world, there is a chance of eventual mutual comprehension. We'll look later at how negotiation can overcome misunderstanding in such circumstances (Ch. 5). For now, note that bodily understanding (that is, understanding in terms of one's body) is only the default case; the rest of the external world can be used to ground one's concepts as—and *if*—needed.

There may, however, be some remaining difficulties in the case of communication between a cognitive agent who uses concepts centered on the human body and one who uses concepts not thus centered. Are there *any* features of the world that are body-*in*dependent (and that, therefore, could serve as a common core for communication)? Lakoff, for one, would probably answer in the negative. For example, colors (as Locke taught us), are human-bodily centered. And one can't even argue that there must at least be something—some primary noumenal feature—in the colored object itself that is the cause of our perception of color. For (as David Zubin has pointed out in discussion) there might be some single color or taste, say, that is perceivable by humans without having a *single* physical cause, but with two or more distinct and non-similar physical causes. Perspective offers a similar example: When I look at an Ames room from a certain angle, I see what you (or even I) would see by looking at a normal room, yet the causes of our similar perceptions are quite distinct.

Still, might there be *some* neutral, objective ways of describing the world that can serve to ground communication? It is difficult to imagine what they would be. Consider as simple a term as 'in'. Lakoff argues that it is human-bodily centered, based on our knowledge of the inside and outside of our bodies (1987: 271–273). But consider a cognitive agent whose body is a "black cloud" in the style of Fred Hoyle's novel (Hoyle 1957). (Such a science-fiction case is necessary here, since the whole point is that if *we* can't imagine how 'in' could be non-objective, then, to imagine it, we need a non-human example. The Black Cloud, not having an inside, might not have a concept of "in". How would such a cognitive agent describe a pea *in* a cup? Topologically, the pea is *on* the cup. So, perhaps, "on" is an objective concept. No matter. I would venture that such remaining objective relations are too few to describe the world. Another example: What about 'inside', as in a pea *inside* a closed box? Perhaps one has no concept of "inside" the box, but the box makes noises if shaken, and, if opened, one sees that now there is a pea *on* the box (in the topological sense of 'on'). Note how hard it would be for the Black Cloud to translate human language (or, at least, English): So perhaps we do need to give our computational cognitive agents *human* concepts.

3.2.2.2.5 Conclusions. A purely syntactic system is ungrounded—it is up in the air, self-contained. But there are arbitrarily many ways to ground it; that is, there are infinitely many possible interpretations for any syntactic system. By "communicational negotiation" (alluded to, briefly, above, and examined further in Ch. 5), we (agree to) ground our language of thought in equivalent ways for all practical purposes. Lakoff and Harnad seek *natural* groundings (cf. "intended interpretations"). Some candidates, such as the (human) body, are convenient. But such natural groundings are merely one or two among many. The only one that is non-arbitrary is the "null" grounding, the "self" grounding: the purely syntactic, "internal" mode of understanding.

3.3 OBJECTIONS.

> ... Turing effectively proposed ... [that] since the question of what rules require (or what formulae mean) is indeterminate, *why not simply build them into the machine itself, which will simply DO what it does and not mean anything at all or be FOLLOWING any rule.* ('Explanations come to a stop' as ... Wittgenstein would put it; 'there is a last house in the lane'.) (Leiber 1991: 54; cf. Wittgenstein 1958: §1, p. 3e, and §29, p. 14e.)

The last semantic domain in a correspondence continuum is the "last house in the lane". It can only be understood syntactically. Hence, all understanding is ultimately syntactic. This is only one of the flaws in John Searle's Chinese-Room Argument. Part of his argument is that computers can never understand natural language because (1) understanding natural language requires (knowledge of) semantics, (2) computers can only do syntax, and (3) syntax is insufficient for semantics. I take part of *my* argument to show that (3) is false (and that, therefore, (2) is misleading, since the kind of syntax that computers do *ipso facto* allows them to do semantics). A few people have disagreed with me. In this section, I'll try to respond to their objections.

3.3.1 Nicolas Goodman's Objections.

In a personal communication (27 January 1987), Nicolas Goodman says that he "can accept ... whole-heartedly" that for me to understand you is for me to provide a semantic interpretation—that (to quote myself) "I map your words into my concepts" (Rapaport 1988: 101). But he would "still not agree with ... [the] conclusion that 'syntax suffices,' since I [Goodman] do not agree that concepts are syntactic objects." What is meant by 'syntactic object' here? Two possibilities come to mind. First, syntactic objects are terms in a formal language. Second, syntactic objects are items in a syntactic domain, where 'syntactic domain' is defined *relative to* a semantic domain; that is, a syntactic domain is that member of a *pair* of domains, one of which (the *semantic* domain) is used to interpret the other (the syntactic domain). In the second sense, as noted above, syntactic objects need not be linguistic ones (that is, need not be syntactic objects in the first sense). In this sense, concepts *are* syntactic objects. I suspect that Goodman objects to treating concepts as syntactic objects in the first sense—the linguistic sense. If so, I can agree with no ill effects, though I would note that insofar as there is a language of thought, it is not unreasonable to take concepts as syntactic objects even in this more specialized sense. Moreover, all that is required for syntax is the existence of rules that characterize the relations among symbols—i.e, some constraints on, or regularities in, their possible behaviors. Without that, *no* domain could be understood.

Goodman offers what I'll call a Japanese-Room Argument: Suppose that Searle-in-the-room can translate, algorithmically, the Chinese input into Japanese squiggles, which he also fails to understand. "However, if we tell him where dinner is served in Chinese, and he translates that information into Japanese, he will still not know anything about how to satisfy his hunger." Of course not! To satisfy his hunger, he would have to do one of two things. First, he could have antecedent understanding of the Japanese. This could either be syntactic understanding—that is, direct understanding of the Japanese—or the ability to interpret Japanese into English. (Note, by the way, that his Japanese translation is a semantic interpretation of the Chinese.) Second, he could translate the Chinese into English (bypassing the Japanese). This was the point of my Korean-Room Argument (Rapaport 1988, §4.1): A Korean Shakespeare-scholar who only reads Shakespeare in Korean translations *does* understand Shakespeare. Similarly, insofar as Searle-in-the-room understands Japanese, he also understands what the Chinese speakers are telling him—in fact, he understands Chinese.

As an aside, let me note that there is, as it happens, some truth to my Korean-Room example, despite the fact that such *indirect* understanding of Shakespeare, or of Chinese, is, of course, *not* what might be called "native" understanding, even though such indirect understanding may sometimes be all that we can achieve:

> During the nineteenth century the vast bulk of Shakespeare scholarship was carried on by Germans, who wrote in German and read Shakespeare (I often suspect—and indeed it was occasionally alleged) chiefly in Schlegel's translations. (Somebody once coined the name of the famous author as August Wilhelm von Schlegelspeare.) Now these German commentators were and are widely respected by native-speaking Shakespeareans. Q.E.D.: these Germans must understand Shakespeare. BUT there are ways and ways of understanding Shakespeare. They have never heard and cannot hear, not understanding English, Shakespeare's *sounds*. They probe his exits and his entrances, his psychological anticipations of Freud, his knowledge of history (defective). But they could do all this without knowing or caring about his music and meter—just as Joseph Papp and other modern producers can put on one of the tragedies and instruct their actors to swallow all the words and speak tripping all over their tongues, so that the poet himself could never recognize any rhythm. Query: does Papp, does the Korean (or German) scholar "know" Shakespeare? Not my Shakespeare—or at best only the fringes of his garment. ...

> Parenthetically: the research of the Yale linguist Helge Kökeritz into the pronunciation of Elizabethan English makes it fairly clear that I too have never heard Shakespeare as he was properly pronounced in his day. Therefore I am just as Korean as the next guy and should not be muddling in these mysterious matters. (Spencer Brown, personal communication, 19 May 1988.)

Moreover, the Japanese-Room Argument doesn't show that concepts aren't, or can't be, syntactic objects. Arguably, the Japanese translations of the Chinese *are* syntactic objects, *not* concepts. And even if they *were* concepts, they would still be in need of interpretation; hence, they would be syntactic in the second sense.

Finally, my semantic interpretation of your syntactic utterances should not be thought of as necessarily a *simple* mapping from one conceptual scheme or semantic network into another. Thus, I can agree with Goodman's observation that "For me to map your words into my concepts ought to mean that I associate with your words various complexes of memory, behavior, affect, etc., in such a way that I end up with a sentence which can play more or less the same role in my life as your sentence plays in your life." I further claim that those complexes *can* all be represented in a semantic network, or perhaps a semantic network linked to input–output transducers, or perhaps a semantic network/input–output complex linked to bodily sensations. The central points are these: (a) All of the things that both Goodman and I are talking about must be linked together. (b) The meaning of any part of such a linked system is (determined by) its location in that whole system. (c) That vast linkage is either understood syntactically, or else by means of an interpretation[12] (in which case, it plays the syntactic role). Thus, I reject Goodman's claim that "Such a mapping would not just involve symbol-manipulations, and so would not be merely syntax."

3.3.2 Neal Jahren's Objections.

Neal Jahren's paper, "Can Semantics Be Syntactic?" (1990), critiquing my theory of syntactic understanding and its application to the Chinese-Room Argument shows how easy it is in discussing these issues to talk just slightly past one another.

What, for example, is a natural language, and what does it mean to understand one? For Jahren, a natural-language is "a series of signs used by a system", and "the sine qua non of natural-language understanding ... [is] an ability to take those signs *to stand for something else ... in the world*" (p. 310, my italics). But if indeed a natural language is just "a series of signs", it follows that to understand it is to understand the series of signs as used by the system—which is a *syntactic* process. Now, as I urged in "Syntactic Semantics" (Rapaport 1988) and in §1.2.6, above, to understand is to map symbols to concepts; thus, for *me* to understand *you* is to map *your* symbols to *my* concepts, which *is*, to use Jahren's phrase, taking "those signs to stand for something else"—but *not* "something in the world" (except in the uninteresting sense that my concepts are things in the world). Moreover, this is still a syntactic process: Insofar as I internalize your symbols and *then* map my internalized representations (or counterparts) of your symbols to *my* concepts, I am doing nothing but internal symbol manipulation (syntax), even though I *am* taking your "signs to stand for something else", namely, my concepts.

Now, do I take *my* concepts to stand for something else outside me? Yes—I so *take* them,

[12] Possibly, a Lakoff–Johnsonian bodily one.

although I only have *indirect* access to the "something else" outside me. The only way I can take *your* symbols "to stand for something in the world" would, pretheoretically, have to be either directly or else indirectly via *my* symbols (concepts). But *all* of it is indirect, since I can at best take your symbols to stand for the same thing I take mine to stand for, and, in both cases, that's just more symbols.

Jahren takes me to task for using 'mentality' in a "suprapsychological" sense (citing Flanagan 1984); he prefers to talk of mentality "in a human sense" (pp. 314ff). But what sense is that? Is it determined by human *behavior* (as in, say, the Turing Test)? If so, then Jahren and I *are* talking about the same thing, since human mental *behavior* might be produced by different processes. Is it determined by the way the human brain does mental processing? But that is too strong for my *computational philosophical* tastes: I am concerned with how mentality, thinking, cognition, understanding—call it what you will—is possible, period. I am not concerned with how *human* mentality in particular works; I take that to be the domain of (computational) cognitive *psychology*. However, I *don't* intend (at least, I don't *think* I intend) the *very* weak claim that as long as a computer can simulate human behavior by *any* means, that would be mentality. I *do* want to rule out table look-up or the (superhuman) ability to solve any mathematical problem, without error, in microseconds. The former is too finite (it can't account for productivity); the latter is too perfect (in fact, if viewed as an infinite, God-like ability to know and do everything instantaneously, it, too, is a kind of table look-up that fails to account for productivity).

Now, having excluded those two extremes, there is still a lot of variety in the middle. So I'll agree with Jahren that, the extreme cases excepted, "a computational system is minded to the extent that the information processing it performs is functionally [that is, input–output, or behaviorally] equivalent to the information processing in a mind" (p. 315)—presumably, a *human* mind. However, Jahren says that two mappings are input–output equivalent "because these mappings themselves can be transformed into one another" (p. 315). This seems to me too restrictive, not to say vague (what does it mean to transform one mapping into another?). Jahren gives as an example "solving a matrix equation [which] is said to be equivalent to solving a system of linear equations" (p. 315). But surely two algorithms with the same input–output behavior would be functionally equivalent even if they were not thus transformable. Consider, for instance, two very different algorithms for computing greatest common divisors. They would be functionally equivalent even if there were no way to map parts of one to parts of the other in any way that preserved functional equivalence of the parts.

Jahren alludes to the symbol-grounding problem: "The semantics$_R$ [that is, the semantics in Rapaport's sense] of a term is given by its position within the entire network" (p. 318). As we've seen, the proper response to this is: 'Yes and no'. *Yes*, in the sense that ultimately all is syntactic. But *no* in the sense that this misleadingly suggests that nothing in the network represents the external world. For instance, Jahren gives an example of 'red' linked as subclass to 'color' and as property to 'apple', etc. But this omits another, crucial—albeit still internal—link: to a `pic`-like node representing the sensation of redness. *Some* parts of the network represent external objects, so an internal analogue of "reference" is possible.

Now, to be fair, Jahren is not unsympathetic to this view:

> ... Rapaport's conception of natural-language understanding does shed some light on how humans work with natural language. For example, my own criterion states that when I use the term 'alligator', I should know that it (qua sign) stands for something

> else, but let us examine the character of my knowledge. The word 'alligator' might be connected in my mind to visual images of alligators, like the ones I saw sunning themselves at the Denver Zoo some years ago. But imagine a case where I have no idea what an alligator is but have been instructed to take a message about an alligator from one friend to another. Now the types of representations to which the word 'alligator' is connected are vastly different in both cases. In the first, I understand 'alligator' to mean the green, toothy beast that was before me; in the second, I understand it to be only something my friends were talking about. But I would submit that the character of the connection is the same: it is only that in the former case there are richer representations of alligators (qua object) for me to connect to the sign 'alligator'. ... The question ... is whether the computer takes the information it stores in the ... [internal semantic network] to stand for something else. (Jahren 1990: 318–319; cf. Rapaport 1988, n. 16.)

Well, the computer does and it doesn't "take the information it stores ... to stand for something else". It *doesn't*, in the sense that it can't directly access that something else. It *does*, in the sense that it assumes that there is an external world—as in our discussion of truth conditions in Chapter 2. But note that if it represents the external world internally, it's doing so via more nodes! There's no escaping our internal, first-person experience of the world (or, as Kant might have put it, there's no escape from phenomena, no direct access to noumena).

I have been avoiding the issue of consciousness and what it "feels like" to understand or to think, though I'll have something to say about *part* of that problem in Chapter 7. But let me make one observation here, in response to Jahren's description of how we can experience what it is like to be the machine: "in accordance with the Thesis of Functional Equivalence one can be the machine in the only theoretically relevant sense if one performs the same information processing that the machine does" (p. 321). That is, to see if a machine that passes the Turing Test is conscious, we would need to *be* the machine, and, to do that, all we have to do is behave as it does. But just "being" the machine (or the "other mind") isn't sufficient—one would also have to simultaneously be oneself, too, in order to compare the two experiences. This seems to be at the core of Searle's Chinese-Room Argument—he *tries* to be himself *and* the computer simultaneously (cf. Cole 1991, Rapaport 1990, Copeland 1993). But he can't use his *own* experiences (or lack of them) to experience his own-qua-computer experiences (or lack of them). That's like *my* sticking a pin into *you* and, failing to feel pain, claiming that *you* don't, either. It is *also* like my *making believe* I'm you, sticking a pin into *me-qua-you*, feeling pain, and concluding that so do *you*. Either one "is" both cognitive agents at the same time, in which case there is no way to distinguish one from the other—the experiences of the one *are* the experiences of the other—or else one is somehow able to separate the two, in which case there is no way for either to know what it is like to be the other. Note, finally, that what holds for me (or Searle) imitating a computer holds for a computer as well: Assume that *we are* conscious, and let a computer simulate us; could the *computer* determine whether *our* consciousness matched *its*? I doubt it.

Let's return to the syntactic understanding of Searle-in-the-room. Jahren says that Searle-in-the-room does not understand Chinese "because ... [he] cannot distinguish between categories. If everything is in Chinese, how is he to know when something is a proper name, when it is a property, or when it is a class or subclass?" (p. 322). I take it that Jahren is concerned with how Searle-in-the-room can decide of a given input expression whether it is a name, or a *noun for* a property, or a *noun for* a class or subclass. In terms of Cassie, this is the question of how she "knows" that 'Lucy' in 'Lucy is rich' is a proper name (how she "decides" whether to build an `object-propername` case frame or some other case frame) or of how she "knows" that 'rich' expresses a property rather than a class (how she "decides" whether to build an `object-propername` case frame rather than a `member-class` case frame).

In one sense, the answer is straightforward: The augmented-transition-network parsing grammar "tells" her. And how does the augmented transition network "know"? Well, of course, we programmed it to know. But in a more realistic case, Cassie would learn her grammar, with some "innate" help, just as we would (see the references cited in §2.8.2). In that case, what the arc labels are is absolutely irrelevant. For us programmers, it's convenient to label them with terms that *we* understand. But Cassie has no access to those labels. So, in another sense, she does *not* know, *de dicto*, whether a term is a proper name or expresses a property rather than a class. Only if there were a *node* labeled 'proper name' and appropriately linked to other nodes in such a way that a dictionary definition of 'proper name' could be inferred (in the manner of §3.2.2.1) would Cassie know *de dicto* the linguistic category of a term. Would she know that something was a proper name in *our* sense of 'proper name'? Only if she had a conversation with us and was able to conclude something like, "Oh—what *you* call a 'proper name', I call a __", where the blank is filled in with the appropriate node label.

Readers who are conversant in reading SNePS networks can get a feel for what it is like to be Cassie by considering the network shown in Figure 3.8 for a Japanese-speaking computational cognitive agent implemented in SNePS (from Arahi & Momouchi 1990: 2). My first reaction on seeing this SNePS network was that indeed I couldn't understand it. But why should I? It only matters for *Cassie* (or her Japanese counterpart) to understand it. *I*, of course, can only understand it by mapping it to my concepts, and there's insufficient information in Figure 3.8 alone for me to do that in any but a non-arbitrary way. In fact, the Japanese networks err in using English arc labels, which makes it *appear* that the arc labels convey some information to Cassie. They don't. They only convey information to *us*; but that's irrelevant.

I'll close this section with a further comment on point 5. Jahren "argue[s] that Searle-in-the-room cannot interpret any of the Chinese terms in the way he understands English terms" (p. 323). But insofar as Searle-in-the-room *is* understanding Chinese, he is *not* understanding English. Neither does Cassie, strictly speaking, understand SNePS networks; rather, she understand natural language, and she uses SNePS networks to do so. But she would only understand SNePS networks if she were a SNePS programmer. And even if she were, the networks she would understand wouldn't be her own—they wouldn't be the ones she was using in order to understand the ones she was programming. Insofar as Searle-in-the-room *does* understand English *while* he is processing Chinese, he could map the Chinese terms onto his English ones, and thus he would understand Chinese in a sense that even Searle-the-author would have to accept.

3.4 SUMMARY.

In this chapter, we have explored the "base" case of our recursive understanding of understanding: the case in which a domain is understood in terms of itself. I have suggested that when a syntactic domain is its own semantic domain, the semantic interpretation function either maps the symbols to themselves or else to other symbols. In the former case, we could only understand the domain by "getting used to it". In the latter case, if there are no fixed points—if each symbol is mapped to a different one (or a set of different ones), then we have the situation we face when using

1．はじめに

自然言語理解システムの構築はマン・マシン・インタフェースおよび電子化された文書データの高度利用という視点から非常に重要な課題である。しかし、自然言語を理解する仕組みの構築には、汎用性や実用性の観点から多くの問題が依然として残されている[1]。この問題に対して、我々は自然言語理解システム構築のための一つの基礎として、日本語文に典型的に出現する名詞述語文の理解について研究を進めている。ここで、名詞述語文とは、主語と述語の対立の中で、述語が名詞で作られる文のことである[2]。名詞述語文については、日本語学の観点から種々の考察が行われている[2,3,4]。しかし、計算言語学的観点から十分研究がなされているとはいえない。このような立場から我々は、従来より名詞述語文「<名詞句１>は<名詞句２>である。」を対象として、名詞句間の意味関係の学習と解析を行う手法の基礎的な考察を行っている[5]。

本稿では、名詞句間の意味関係を学習・解析する手法の概要とその有効性を評価するため本手法に基づく実験システムを作成して行った実験結果について述べる。

2．名詞句間の意味関係

名詞述語文における<名詞句１>と<名詞句２>の間の基本的な意味関係を以下に示す。

（１）下位・上位関係
<名詞句２>が<名詞句１>の上位概念を表わす。
例：鯨はほ乳類です。

（２）同一関係
<名詞句１>と<名詞句２>が同一概念を表わす。対象レベルと概念レベルでの同一関係がある。
例１：富士山は日本一の山です。
（対象レベルでの同一関係）
例２：正方形は直角正四辺形です。
（概念レベルでの同一関係）

（３）対象・属性関係
<名詞句１>が対象で<名詞句２>が<名詞句１>の属性値を表わす。
例：太郎は医者です。

（４）対象・事象関係
<名詞句１>が対象で<名詞句２>が<名詞句１>の関与する事象を表わす。
例：父が帰宅です。

（５）要素・集合関係
<名詞句１>が<名詞句２>の表わす集合の要素となっている。
例：雄は性別です。

（６）うなぎ文関係[6]
「太郎は鰻です。」に代表される文で文脈に依して、例えば「太郎は鰻を食べます。」という意味に解釈される。これは、太郎という名前の鰻がいて文字どうりそれが鰻であるという解釈も可能である。

（７）比喩関係
<名詞句１>の属性に<名詞句２>の顕著な属性を重ねる隠喩を表わす。
例：太郎は鰻です。
（太郎は鰻のようにぬらりくらりと言い訳をする。）

（８）同語反復同一関係
<名詞句１>と<名詞句２>に同じ名詞句を置き、文脈（状況）の中で修辞的解釈が行なわれる。
例：太郎は太郎です。自分で考え、自分で行動すべきです。

ここでは、文字どおりの意味である（１）～（５）を対象とし、意味関係の学習と解析を行う。文字どおり以外の意味である（６）～（８）についての学習と解析については、今後研究を進め別の機会に報告したい。

3．知識表現

我々は、知識表現として、S.C.Shapiroらによって提案された意味ネットワークSNePS[7,8]に基づく表現を用いている。図１に「人間はほ乳類です。」のSNePS表現の例を示す。SNePSでは節（nodes ）が、命題、対象、属性、関係などを表わし、弧（arcs）はそれらの間の構造的、意味的つながりを表わす。ここで、ある概念からある概念までの弧の連なりを二つの概念間のパス（経路）と呼ぶ。図１で<lex>につながる<人間>と<ほ乳類>は語彙であり、<m１>と<m２>は<人間>と<ほ乳類>の概念であり、<m３>は<m１>と<m２>が下位・上位関係であるという概念である。

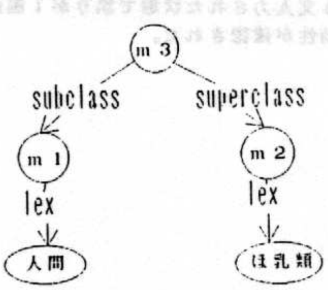


図１　SNePSの例

ここで用いる概念生成モデルは、事物の概念である具象概念を事物の性質に対応する属性概念によって制限することにより新たな概念である派生概念を生じ、それが具象概念あるいは属性概念として知識の中に組み込まれてゆくというものである。したがって、概念は具象概念、属性概念、派生概念のいずれかに分類される。これは、学習の段階でヒューリスティックスを用いて行われる。

ここで、２章で述べた（１）～（５）までの意味関係をネットワーク上での弧で表現する際の表記について述べる。ここで、（１）、（３）、（４）、（５）については、それぞれ<名詞句１>を始点とするか<名詞句２>を始点とするかで２通り考えられるので、実際には以下のような９つの関係が考えられる。

（１）下位・上位関係：<-subclass-:-superclass->
　　　上位・下位関係：<-superclass-:-subclass->
（２）同一関係　　　：<-equiv-:-equiv->
（３）対象・属性関係：<-object-:-property->
　　　属性・対象関係：<-property-:-object->
（４）対象・事象関係：<-object-:-event->
　　　事象・対象関係：<-event-:-object->
（５）要素・集合関係：<-member-:-class->
　　　集合・要素関係：<-class-:-member->

4．概要

Figure 3.8: A Japanese-speaking computational cognitive agent.

a dictionary. The difference is that since all external items are also mapped into internal ones, the symbol-grounding problem can be avoided. If there *are* fixed points, then they come to be understood either retroactively in terms of the role they play in the understanding of other terms, or else by "grounding" them to "non-linguistic"—albeit *internal*—symbols.

In any case, we have a closed network of meaning—a "conceptual-role semantics". We explore the implications of this in the next chapter.

Chapter 4

CONCEPTUAL-ROLE SEMANTICS

> Who knows only one thing knows not even that. A thing entirely isolated would be unknowable. There would be nothing to say of it, or any language for it. The reason for this has been obvious to people as different as Saint Thomas Aquinas and William James. Thomas said: "the soul is pleased by the comparison of one thing with another, since placing one thing in conjunction with another has an innate affinity with the way the mind acts." [Saint Thomas Aquinas, *Summa Theologiae* I-II 32: 8.] And James said: "the first thing the intellect does with an object is to class it along with something else." [William James, *The Varieties of Religious Experience*, Lecture 1]. (Wills 1991: 18.)

> The question of how one understands the language one thinks in does seem to be a peculiar one. ... C[onceptual]R[ole]S[emantics] clarifies the situation. (Loewer 1982: 310.)

4.1 CONCEPTUAL-ROLE SEMANTICS AND HOLISM.

In §2.4, I talked of pattern matching as the way to determine correspondences between two domains. When two patterns A and B match, the result is a determination that a part of pattern A "plays the same role" in pattern A that a corresponding part of pattern B plays in pattern A. That role, I suggested, was the part's syntactic role in its own domain—a role determined by the part's internal relationships to the other parts of the pattern. I argued in Chapter 3 that this is where semantics "bottoms out", in the syntactic understanding of a (syntactically specified) domain, where what counts for a term's meaning is its syntactic role. This kind of semantics has come to be called "conceptual-role semantics" or "inferential-role semantics". In this chapter, we'll look at two major conceptual-role semantic theories and reply to several objections to them.

Conceptual-role semantic theories are almost always associated with holistic theories of semantics. Both have lately come under sustained attack from Jerry Fodor and Ernest Lepore (*Holism*, 1992), who argue that there are no good arguments for holism. That may be, yet I find holism attractive. I take my task in this chapter not so much to argue *for* it (I doubt that I could find an argument stronger than those refuted by Fodor and Lepore) as to paint an attractive picture of holism and conceptual-role semantics and to clarify that picture in the light of the critiques of holism and conceptual-role semantics.

As before, the model I'll be thinking of is a vast semantic network—a propositional, intensional semantic network with ways of incorporating sensory images among its nodes; in short, SNePS. But this is merely to fix ideas; nothing else hinges on it.

The quotation from Gary Wills that opened this chapter nicely expresses the core ideas behind holism and conceptual-role semantics. Once cannot say anything about an isolated node without thereby enlarging the network and de-isolating the node. As such a process continues, the network grows. This is how holistic conceptual-role semantics begins. Since all that is initially known about the isolated node is now expressed in the rest of the network, the node's "meaning" is determined by its location or role in that entire network (Quillian 1967, 1968). Nodes that are very distant from the original one may have little to do directly with its meaning or role. But they will have something to do with other nodes that, eventually, directly impact on—or are impacted on by—that original node. To use an old terminology, they may be part of that node's "connotations". (Hill 1994 provides a formal interpretation of this.)

The larger the network, the more meaning its nodes have—that is, the more can be said about them—and the larger their roles are. Stephen Stich (1983) has argued that a person with a single, isolated "belief" does not really have any beliefs. I would prefer to say that the more beliefs one has, the more each belief means. Such an isolated belief *is* a belief, but not one that has much of a role to play. (Similarly, as I pointed out in "Syntactic Semantics" (1988), linguists who build syntactic and semantic theories from studies of isolated sentences would also do better to look at connected discourse.)

Isolation—even a complex network that is isolated from the rest of the network—is a barrier to comprehension. A patient can convey, without understanding it, a message from a doctor to a dentist, both of whom *will* understand it (cf. Rapaport 1988: 126n16), because the medically ignorant patient cannot link the message to his or her own semantic network, while the medical personnel can link it to theirs. Consider a fax machine. It takes text, converts it to electronic signals, and reconverts these to text. Yet—like the patient—it has no "knowledge" of the text. We seem to have a Chinese Room. But if the conversion were, say, to ASCII code, which could be linked to a knowledge base, we might have an "intelligent" fax machine. It is the *links* that count; isolation doesn't yield understanding:

> In most cases it is not possible to infer the meaning ascribed to a symbol within a given culture from the symbolic form alone. At the very least, we have to se how that form is used, how it is reacted to. We have to see it in the context of other actions and of other speakers. (Renfrew 1990: 7.)

It is always, of course, a matter of degree. If "an elephant is so he can have a trunk" (Spencer Brown, personal communication), and that's all we know about elephants or their trunks, then all we know about their trunks is that they can be had by elephants. But as our knowledge of elephants (and their trunks) enlarges, we come to understand more (and, no doubt, to express it more informatively, less obviously circularly):

> [T]he problem of 'genuine semantics' ... gets easier, not harder, as the K[nowledge]B[ase] grows. In the case of an enormous KB, such asCYC's, for example, we could rename all the frames and predicates as G001, G002, ..., and—using our knowledge of the world—reconstruct what each of their names must be. (Lenat & Feigenbaum 1991: 236.)

Carnap said as much—years earlier—in his example of a railroad map. There, he showed how to describe any object in a given domain in terms of the other objects, without any external "grounding" (Carnap 1928, §14, pp. 25–27; cf. Rapaport 1988: 111).

But note some potential problems. The network can't be too *simple*, for then it would be underspecified (cf. Rapaport 1988: 123–124). It would be a pattern that was too general, that would match too much. But neither can the network be too *complex* (as in the case of CYC): Although a giant pattern-matching procedure as envisaged by Lenat and Feigenbaum (1991) is possible in principle, I don't see how it could be carried out in practice very easily. Better to let the nodes (some of them, at least) wear their intended interpretations on their sleeves. To switch examples back to SNePS, it is better to let a `LEX`-node labeled 'rich' be expressed by the English word 'rich' than by something arbitrary. (Even this might not be needed if smaller, more tractable portions of the full knowledge base could be understood in the manner that Lenat and Feigenbaum suggest.) This is what we do when we talk to each other. More on that later (Ch. 5).

Let's now look at two of the major conceptual-role semantic theories, the early, influential one of Wilfrid Sellars and the more recent one of Gilbert Harman.

4.2 SELLARS'S THEORY OF LANGUAGE GAMES.

> As I see it, abstract singular terms such as 'redness' ... and 'that Chicago is large' are to be construed, in first approximation, as singular terms for players of linguistic roles .... (Sellars 1961/1963: 204.)

4.2.1 Cassie.

In a series of papers that became chapters of his *Science, Perception and Reality* (1963), Wilfrid Sellars spelled out a classic theory of conceptual-role semantics.[1] Before commenting on it, it will be useful to think in terms of SNePS/Cassie.

In "The Language of Theories" (1959/1963: 109–113, §§11–18), Sellars distinguishes a variety of kinds of meaning:

[1] "The Language of Theories" (1959/1963), "Truth and 'Correspondence'" (1961/1963), and, especially, "Some Reflections on Language Games" (1955/1963).

meaning as translation:
'round' means *circular*;[2]
'cheval' means *horse*.

meaning as sense:
'round' expresses the concept Circularity;[3]
'cheval' expresses the concept Horsekind.

meaning as naming:
'round' names the concept Circularity;[4]
'cheval' names Man O'War.

meaning as connotation:
'cheval' connotes the property of having four legs;
'Parigi' connotes the property of being the capital of France.

meaning as denotation:
'round' denotes circular things.[5]

Conceptual-role semantics is about meaning as translation, though there is room for all the others (except possibly the last—but see Ch. 3).

What about Cassie? Suppose she hears Oscar say that something "is round". Insofar as Cassie maps Oscar's utterance or use of 'round' to her own 'round' node, she is understanding Oscar by translating his utterances into her semantic network. (If she has never heard 'round' before, she'll create a new node on which to map Oscar's utterance; it's still translation.) I would say, however, that Cassie's `LEX` node labeled 'round' expresses the *concept* at the tail of the `LEX` node. Thus, in Figure 4.1, node `M1` is Cassie's concept of roundness (or circularity, to use Sellars's somewhat misleading locution). If Cassie wanted to talk about that node (and to say more than that something (viz., `B1`) is round), she could name it; node `M3` would be its name, expressed as 'Circularity'. (Here, I differ a bit from Sellars.) Connotation can be accounted for, in part, as follows: Suppose Cassie learns that round things have curved surfaces (Figure 4.2, node `M5`). Here, the connotation of 'round' is given (in part) by rule node `M5` (as well as, perhaps, by `M2` and `M4`, and so on, throughout the full network).

Denoting, however, is a relation that Cassie cannot deal with for herself. It is an external, third-person relation. However, Oscar could assert that Cassie's 'round' denotes some round thing. We have the situation shown in Figure 4.3. According to Sellars, Cassie's word 'round$_C$' denotes some circular thing, α; so denotation, for Sellars, is a relation between a word and an external object. As such, it is not accessible to Cassie. (By the way, presumably there are also relations, equally inaccessible to Cassie, between her *concept* of roundness, viz., $\texttt{M1}_C$, and α, and between her concept of α, viz., $\texttt{B1}_C$, and α.) From Oscar's point of view (not much different from *our* point of view with respect to Cassie), Cassie believes that something (which Oscar represents as $\texttt{B2}_O$) is round, and Oscar can believe that Cassie's word 'round' (actually, *Oscar's* representation of her word) denotes (in Sellars's sense) the object (that Oscar believes) that Cassie believes is round, viz.,

[2]I would prefer to say that 'round' means *round*.
[3]I would prefer to say that 'round' expresses the concept Roundness.
[4]I would prefer to say that 'round' names the concept Roundness.
[5]I would prefer to say that 'round' denotes round things.

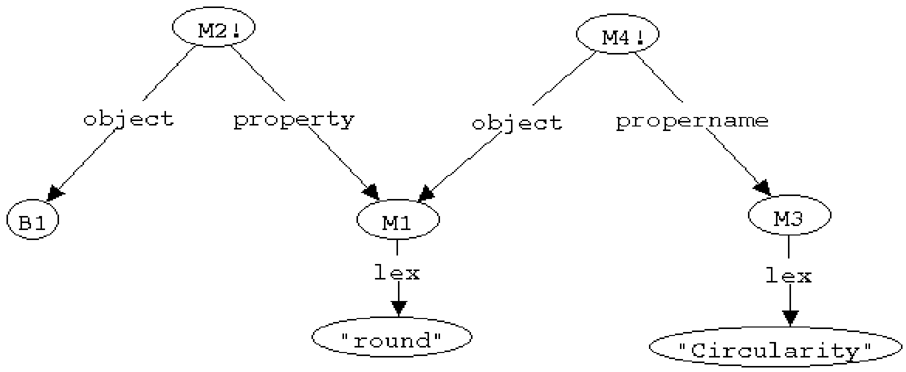


Figure 4.1: Something is round and the concept of roundness is named 'Circularity'. (`M2!` = `B1` is round; `M4!` = `M1` is named 'Circularity'.)

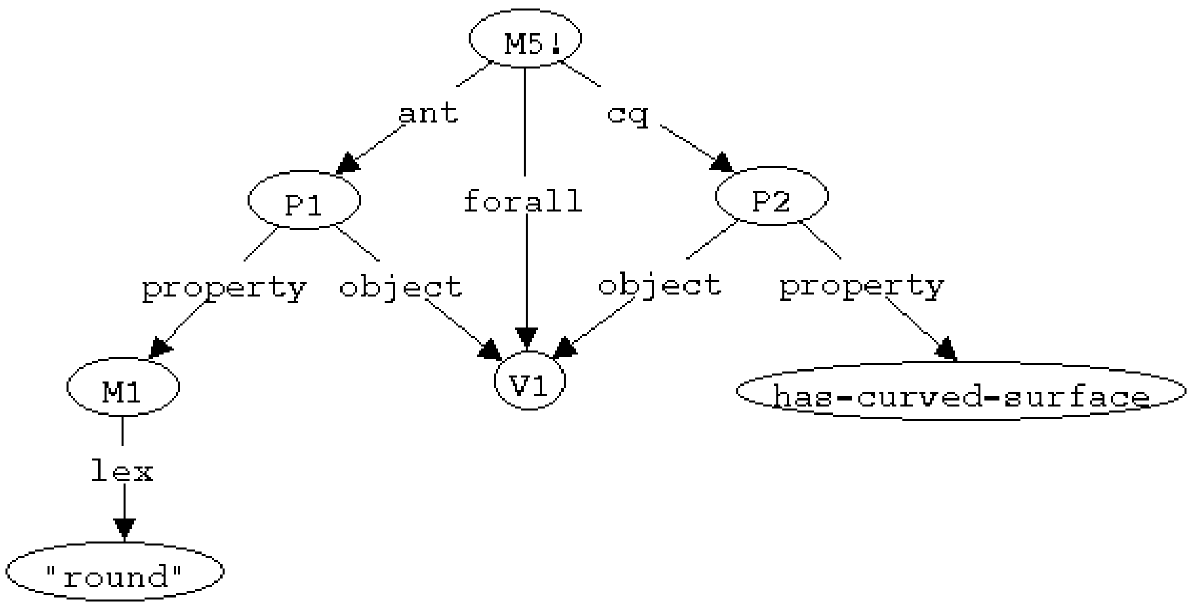


Figure 4.2: Round things have curved surfaces. `M5!` = $\forall$`V1`[Round(`V1`) $\rightarrow$ Has-Curved-Surface(`V1`)], where, for the sake of the example, 'Has-Curved-Surface' is not—but could be—further analyzed. (Node `M1` here is the *same* node as node `M1` in Figure 4.1.)

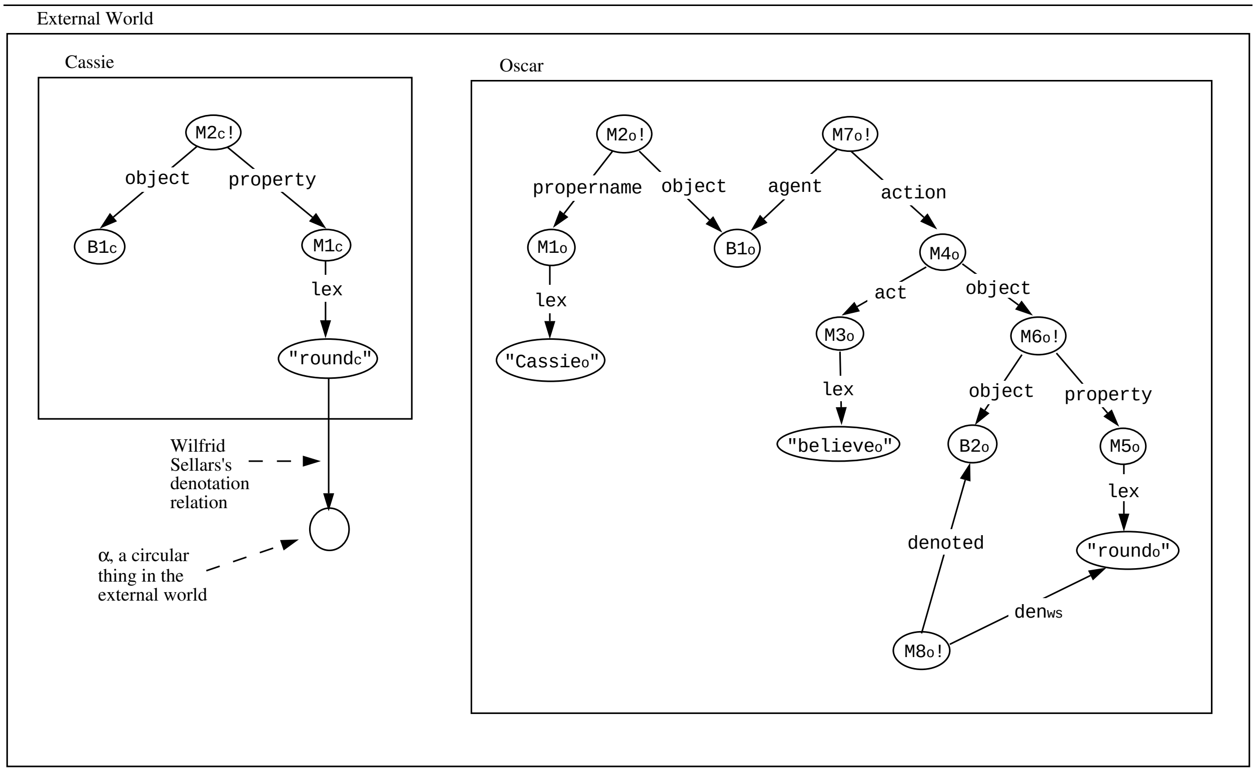


Figure 4.3: Cassie's and Oscar's representations that something is round.
In the external world, Cassie's node `"round"` denotes-in-Sellars's-sense α.
In Cassie's belief space, $\texttt{M2}_C\texttt{!}$ $= \texttt{B1}_C$ is round.
In Oscar's belief space, $\texttt{M2}_O\texttt{!}$ $= \texttt{B!}_O$ is named 'Cassie$_O$',
$\texttt{M7}_O\texttt{!}$ $= \texttt{B1}_O$ believes that $\texttt{M6}_O\texttt{!}$,
$\texttt{M6}_O\texttt{!}$ $= \texttt{B2}_O$ is round,
$\texttt{M8}_O\texttt{!}$ $=$ 'round' denotes-in-Sellars's-sense $\texttt{B2}_O$

$\mathtt{B2}_O$. (Again, presumably, there are relations, equally inaccessible to Oscar, between the following pairs: 'round$_O$'/'round$_C$', $\mathtt{B2}_O/\mathtt{B1}_C$, $\mathtt{B2}_O/\alpha$, $\mathtt{M5}_O/\mathtt{M1}_C$, and $\mathtt{M6}_O/\mathtt{M2}_C$.)

What can we say about statements like the following?

1. $\ulcorner x \urcorner$ means y',
2. $\ulcorner \varphi \urcorner$ is true',
3. $\ulcorner x \urcorner$ is about y'

I'd say first that they're missing a parameter or two. The statements should really be, respectively:

1. Cognitive agent C's use of $\ulcorner x \urcorner$ means y for cognitive agent O;
2. cognitive agent O believes that cognitive agent C's utterance or belief that $\ulcorner \varphi \urcorner$ is true;
3. cognitive agent C's use of $\ulcorner x \urcorner$ is about what cognitive agent O refers to as $\ulcorner y \urcorner$.

So, let me answer the question from Oscar's point of view. (1) For Oscar to say that Cassie's use of $\ulcorner x \urcorner$ means y is to say that Cassie's use of $\ulcorner x \urcorner$ plays the same role in *her* belief system that $\ulcorner y \urcorner$ plays in his (Oscar's). (2) For Oscar to say that Cassie's utterance of $\ulcorner \varphi \urcorner$ is true is to say that he endorses her utterance of $\ulcorner \varphi \urcorner$; that is, it is to say that he believes it (too); cf. Rapaport, Shapiro, & Wiebe 1986. As Sellars puts it,

> In general, when I commit myself to
>
> (w) S is a true sentence (of L)
>
> I am committing myself to asserting either S itself (if I am a user of L) or a translation of S into the language I do use. (Sellars 1955/1963: 354, §78.)

(3) For Oscar to say that $\ulcorner x \urcorner$ is about y is for him to say that he interprets $\ulcorner x \urcorner$ by y, where both are nodes in *his* network: $\ulcorner x \urcorner$ is a syntactic, or linguistic, node; y is a semantic, or non-linguistic, node. If Oscar wants to say what his *own* word $\ulcorner x \urcorner$ means, he must do it in that way, too: asserting a link between it and some other fragment of his network.

4.2.2 Reflections on "Reflections on Language Games".

Sellars's essay "Reflections on Language Games" (1955/1963) is relevant to several of our concerns, and we'll return to it in Chapter 7, when we look at *implementation*. Here, I want to concentrate on the syntactic nature of his conceptual-role semantic theory.

For Sellars, to use a language is to do certain actions in certain circumstances—presumably, for example, to utter certain expressions in certain circumstances—and this is to be viewed as making "moves" in a "language game" whose "positions" are "awareness" "of propositions, properties, demands, **[???]** etc." (pp. 324, 327, §§10, 16). There are three kinds of such moves (p. 328 §§19–23):

1. "language-entry transitions" from observations of the external world to positions in the language game (that is, input, in which the position "means" the observation; cf. p. 329, §22);

2. "moves", or inferences, between positions in the language game (that is, relations among sentences);

3. "language-departure transitions" from "ought-to-do" positions to actions (that is, output, in which the position "means" the action) (cf. p. 329, §23).

In terms of Cassie, language-entry transitions occur when she finds or builds a node in her semantic network as a result of something she hears, reads, or perceives, and language-departure transitions occur when she utters something as a result of an intention to speak or when she performs an action as a result of an intention to act (cf. Bruce 1975; Cohen & Perrault 1979; Allen & Perrault 1980; Cohen & Levesque 1985, 1990; Grosz & Sidner 1986; Haller 1993abc; Kumar 1993abc, 1994ab; Kumar & Shapiro 1993). The internal, inferential moves correspond to *any and all internal processing* of the semantic network. They need not all be "inference" in any strict logical sense (hence my preference for the term '*conceptual*-role semantics'). Of course, the input positions could be established in other ways (e.g, by direct manipulation by a "computational neuroscientist", or Wilder-Penfield–like stimulation). For instance, Sellars also allows "free" positions: sentences that are neither the result of internal, inferential moves nor of observations—roughly, they would be axioms or "primitive" beliefs: sentences taken on faith (p. 330, §25). And the output positions need not result in (successful) action (as long as the system *believes* that it does—cf. the blocks-world robot of §2.7.1).

To thus "speak of a language as a game with pieces, positions, and moves" is to treat it purely syntactically. "But must we not at some stage recognize that the 'positions' in a language *have meaning* … ?" (p. 332, §30). This is the key issue. Note, however, that for Sellars it would not be the *pieces* that are to "have meaning", but the *positions*. The quotation from Sellars 1961/1963 that opened Section 4.2 should be taken literally. In Figure 4.1, the *term* 'Circularity' is a proper name for a concept, viz., `M1`, and it is the concept that *is* the role. What *plays* the role is the term 'round'. Strictly speaking, then, we could say that, for Cassie, 'round' means node `M1`, whose *role* is specified by its location in the network.

According to Sellars,

> … the German expression 'Es regnet' … *means* it is raining. … [I]n saying this …, one is not saying that the pattern 'Es regnet' plays a certain role in the pattern governed behaviour to be found behind the Rhine. But it would be a mistake to infer from these facts that the semantical statement ' "es regnet" means *it is raining*' gives information about the German use of 'Es regnet' which would *supplement* a description of the role it plays in the German language game, making a *complete* description of what could otherwise be a partial account of the properties and relations of 'Es regnet' as a meaningful German word. (p. 332, §31.)

I interpret this as a negative answer to the question whether "positions" have "meaning"—at least, a "meaning" as an *external* item. For Sellars, it seems, syntax suffices. Although there is a non-syntactic, externally-semantic dimension to meaning, it has nothing to do with the language game. Cassie's (internal) ability to use language is syntactic; Searle's Chinese-Room Argument is wrong.

Sellars continues,

> To say that ' "*rot* means *red*' is not to describe 'rot' as standing 'in the meaning relation' to an entity *red*; .... (p. 332, §31.)

That is, semantics is *not* a correspondence between language and the world.

> ... it is to use ... the semantical language game ... for bringing home to a *user* of 'red' how Germans use 'rot'. (p. 332, §31.)

That is, semantics *is* a correspondence between two *languages*: between the speaker's language and the third-person, external observer's language (and perhaps that observer's concepts, too). English-speakers understand a German-speaker's use of 'rot' as their (i.e., the English-speakers') concept *red* (as the concept they express with 'red'). This is semantics in the classic sense: The English-speaker uses a model for interpreting the German-speaker's utterances. But the model is just the English-speaker's own language game—a syntactic system.

> To say that ' "*rot*" means *red*' ... conveys no information which could not be formulated in terms of the pieces, positions, moves, and transitions (entry and departure) of the German language game. (p. 332, §31.)

Again, it's purely syntactic.

(Actually, I think Sellars should have said '*English* language game' here: I'm assuming that the English speaker wonders what 'rot' means and is told that it means *red*. The English speaker now has nodes representing the German word 'rot' and the concept it expresses; and the English-speaker maps these—internally—to the nodes representing the English word 'red' and the concept *it* expresses. Thus, all of the information conveyed by the 'rot'-means-*red* sentence can "be formulated in terms of the pieces, positions, moves, and transitions ... of the *English* language game".)

Sellars discusses a cousin of the symbol-grounding problem under the rubric "prelinguistic concepts":

> Now there appear to be two possible lines that can be taken with respect to such ur-concepts:
>
> (1) They are interpreted as a structure of symbols and, hence, *in our broader sense*, as a *language*. ... [A] regress is lurking which can be stopped only by admitting that the meaningfulness of at least one symbolic system is not clarified by the idea of obeying semantical rules.
>
> (2) As a second alternative, the ur-concepts may be conceived as pre-symbolic abilities to recognize items as belonging to *kinds* .... (pp. 334–335, §37.)

Possibility (1) is my purely syntactic view, and I do "admit" "the meaningfulness of ... [a] symbolic system [that] is not clarified by the idea of obeying semantical rules" such as "red objects are to be called 'red'" (Sellars 1955/1963: 334?, §37). To clarify the "meaningfulness" of such a symbolic system, we need internal—syntactic—understanding. Possibility (2) is the Harnad–Lakoff alternative, which Sellars rejects on the grounds that it commits the homuncular fallacy.

Sellars urges a distinction between 'bishop' in chess and 'piece of wood of such and such shape' (p. 343, §56), and he then elaborates on possibility (1):

> . . . I might learn to respond to the move-enjoining sentence 'Sellars, advance your king's pawn!' as I would to 'Sellars, shove this piece of wood two squares forward!' (p. 344, §57.)

Compare the Chinese Room: "shoving a piece of wood forward" is the rule-book's translation of the meaningless squiggle "advance your king's pawn". Perhaps, though, shoving that piece forward just *is* advancing one's pawn in the same way that talking of certain chemical structures just *is* talking of mathematical lattices. We'll make this sense of "is" more precise in Chapter 7. In any event, Sellars rejects it:

> But while this *might* be the description of learning to apply the rule language game . . . , it would make the connection between expressions such as 'bishop' . . . in chess language and the expressions in everyday language which we use to describe pieces of wood, shapes, sizes, and arrangements much more 'external' than we think it to be. For surely it is more plausible to suppose that the piece, position, and move words of chess are, in the process of learning chess language, built on to everyday language by *moves* relating, for example, 'x is a bishop' to 'x is a ♗-shaped piece of wood' . . . . In other words, chess words gain 'descriptive meaning' by virtue of *syntactical relations* to 'everyday' words. (p. 344, §58.)

As I have urged with respect to the Chinese-Room Argument, pulling the semantic rabbit out of the syntactic hat is no trick—it's all done with *internal* links. My understanding of 'bishop' (or 'pawn', or Searle-in-the-room's understanding of a Chinese squiggle) is *not* provided by an external link to a ♗-shaped piece of wood, but by an *internal*, *syntactic* link to my internal representation of such a ♗-shaped piece of wood.

The fundamental thesis of conceptual-role semantics, as formulated by Sellars, is that

> statements of the form
>
> '. . . ' means — (in L)
>
> are incorrectly assimilated to relation statements. . . . [Rather,] they convey . . . the information that '. . . ' plays the role in L which '—' plays in the language in which the semantical statement occurs. (pp. 354–355, §80.)

Of course, if the semantic language *is* L, the meaning of '. . . ' would have to be given in terms of the role it plays in L, by specifying its location in the network—its position in the game.

Let's now have a look at Harman's variations on Sellars's theme.

4.3 HARMAN'S THEORY OF CONCEPTUAL-ROLE SEMANTICS.

In a series of papers, Gilbert Harman has advocated a Sellarsian conceptual-role semantic theory *almost* all of which is congenial to the view taken here (Harman 1974, 1975, 1982, 1987, 1988, and esp. 144: 283–284). **[???] Gunderson, Lang, Thought, Commn?** The issue can be approached by asking whether an internal conceptual-role semantics based on *translating* one language into another is all that is needed to explain our knowledge of the semantics of language, or whether an external referential/truth-conditional theory plays a role (if you'll excuse the expression) (Harman 1974: 1). I called the latter kind of theory 'external', but it is actually both internal and external; that is, it must be a bridge theory that links an internal syntactic domain with an external semantic domain. *Perhaps* such a theory could tell us something about the denotations of terms and the truth values of sentences. But, of course, since the cognitive agent has no access to the denotations or states of affairs themselves, a theory of truth tells the *agent* nothing. It is simply not available to the agent, who is restricted to the internal point of view. Now, as Harman notes, theories of truth do shed light on meaning—consider possible-worlds model-theoretic semantics for modal logics, clearly a major intellectual achievement. But note, first, that such theories are addressed to professional philosophers and cognitive scientists, who are external observers: Oscar can use such a theory to understand the relation of Cassie's language to the world, but he doesn't use the theory when he understands Cassie in everyday conversation. Second, as we learned from Smith's Gap, truth theories are correspondences between language and a *model*, not between language and *the world*. So they themselves are translations—between the language playing the syntactic role and the language of the model.

There are two other possible roles for truth theories or external links. One, relevant to Sellars's "entry" and "departure" rules, we'll come back to shortly. The other is the role of truth in logical inference, Sellars's internal "moves": "logical implication is a matter of truth and logical form" (Harman 1974: 11). But here, truth is only a sort of place holder: Logical implication must preserve truth, but no claims are ever made about actual truth *values*, nor need they be. The rules of inference of a syntactic system are themselves purely syntactic, as we saw in Chapter 2. They need not—indeed, *do* not—mention truth. In a given system, some rules might be preferable to others (they can be justified) because they *preserve truth*. That plays a role in which rules to choose, but not in the actual working of the rules. Indeed, that's the whole point of syntactic systems: We devise them in order to talk about truth, so we want them to *represent* truths. The world, together with its objects, relations, states of affairs, and truths, is one thing; the language, with its corresponding terms, relation symbols, wffs, and rules of inference and theorems used to discuss the world, is another. We want language and the world to *correspond*; they don't intersect.[6] From the internal, first-person point of view, all that we *can* deal with is the syntactic theory. And, if all we're dealing with is the syntactic theory, we don't need truth at all. Or, rather, *Cassie* doesn't need it, and can't have it anyway, and *Oscar* (who studies Cassie's language-use from the external, third-person point of view) has access to truth only as a correspondence among *beliefs* (Harman 1974: 9): Oscar translates Cassie's utterances into his own semantic network. If he tries to say what *is* true, all he can do is to say what *he* believes: If he didn't believe it, he wouldn't try to claim that it's true. That is, for Oscar to say that φ is true is just for him to say that (he

[6] Well, actually they do, of course: The language is part of the world. But that fact is ignored when the language is used to describe (the rest of, or some other part of) the world. Cf. the description of Figures 1(I) and 1(II) in Rapaport 1985/1986: 67–71.

believes that) φ. For Oscar to say that what Cassie said is true is also just for him to say that he believes what Cassie said (cf. Rapaport, Shapiro, & Wiebe 1986; Roberts & Rapaport 1988).

How do truth conditions provide the meaning of a sentence? 'Snow is white' is true if and only if snow is white; so, 'snow is white' *means* that snow is white. There are two well-known problems with this. First, 'snow is white' is also true if and only if grass is green (at least, this would be so when snow is white if and only if grass is green), but 'snow is white' doesn't *mean* that grass is green.[7] Second, although 'All mimsy were the borogoves' is true if and only if all mimsy were the borogoves, to say that 'All mimsy were the borogoves' *means* that all mimsy were the borogoves clarifies little (Harman 1974: 6; it's the circular dictionary problem, with a circle of radius 0). What's missing is knowledge of what 'mimsy' and 'borogove' mean. How could we find out? We could find the denotations, but that's solipsistically impossible. Alternatively, we could find our mental representations (of the denotations) (cf. Harman 1974: 6), or we could give a definition of the terms: Both of these are purely internal and syntactic, however.[8]

Consider both the white-snow and the mimsy-borogoves cases from Cassie's point of view. She hears 'snow is white', and she understands it by mapping 'snow' onto her concept of snow, 'white' onto her concept of white, and forming the proposition that snow is white. That is, she understands the sentence by forming that proposition, which is now linked to her semantic network. She *believes* that snow is white if and only if either she already had a mental representation of that proposition ("Oh yes; I already knew that") or she has reason to trust the speaker ("Oh yes? Well, if you say so"). If she hears 'all mimsy were the borogoves', she will seek to understand by finding (or building) a mimsy-node and a borogove-node, and finding (or building) the proposition that the borogoves were entirely mimsy. But she won't understand it as well as she understands the proposition that snow is white, since it will *not* be linked to the rest of her network. (Or, at most, it will be linked to her representation of the rest of *Jabberwocky*. So, at best, she'll have a skeletal understanding in the context of the poem.)[9]

It may be objected that this is an example from literature, so talk of truth conditions is beside the point. But, as Harman points out, that's *part* of the point: "Speakers violate no linguistic conventions when they ... tell stories" (Harman 1974: 10). So it is not the case that we must claim that speakers try to say what's true. Rather, at most we only have to claim that they try to say what they *believe*. But they don't even always try to do *that*: Sentences from fiction are, depending on your tastes, either false, truth-valueless, or the sort of thing for which a truth theory would be a category mistake (cf. Ryle 1933; Parsons 1975; Searle 1979; Pavel 1986; Castañeda 1979, 1989a; Rapaport 1991a; Rapaport & Shapiro 1995.) In any case, a truth theory yields strange results when applied to sentences from fiction (though no stranger, perhaps, than when applied to modal sentences that require possible—if not fictional—worlds).

The point is that semantics as correspondence between language and *the world* is of no help in giving a first-person explanation of how a cognitive agent understands language. (And it is certainly of no help in giving a first-person explanation of how a cognitive agent understands

[7]Although, when it snowed on the first day of Spring the year that I wrote this, I cheered myself up by thinking so!

[8]Or we could define one in terms of the other, as suggested above in §§3.2.2.1, 4.1: Borogoves are things that can be mimsy, or else being mimsy is something that borogoves can be. Again, this tells us little by itself (more context is needed). In any case, it is still purely syntactic.

[9]Or it may be linked to her representations of the rest of *Through the Looking Glass*, in which Humpty Dumpty explains the sentence. In that case, she'll understand it, because further links will have been made. The more links, the more understanding.

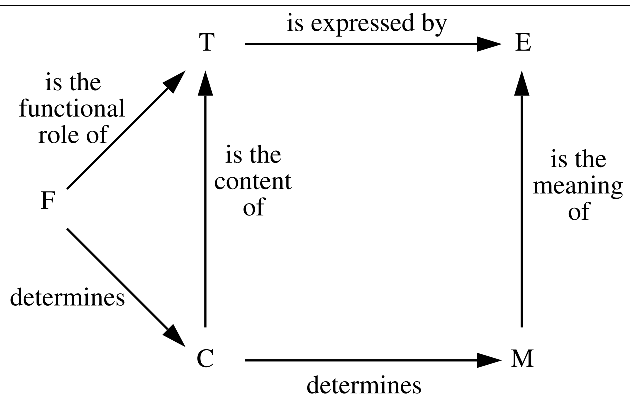


Figure 4.4: The meaning, M, of a linguistic expression, E, is determined by the content, C, of the thought, T, that is represented by E; the functional role, F, of T determines C.

fictional language.) However, semantics as correspondence between language and the agent's mental representations (or language of thought) *can* help: "The meaning of a sentence is determined by the thought with which the sentence is conventionally correlated, that is, the thought which, by convention, speakers would normally intend to communicate to a hearer by using that sentence" (Harman 1974: 10). Of course, to talk of "the" meaning of a sentence is misleading. Context needs to be taken into account. But the broader point holds: Meanings of sentences are provided by thoughts, not by truth conditions.

Harman, however, formulates this a bit differently from the way that I see it: There are three parts to his formalism. Here are the first two:

1. The meanings of linguistic expressions are determined by the contents of the concepts and thoughts they can be used to express. (Harman 1982: 242; 1987: 55.)

2. The contents of concepts and thoughts are determined by their functional role in a person's psychology. (Harman 1982: 242.)

And, in a later essay, part 2 is analyzed further:

2a. The contents of thoughts are determined by their construction out of concepts. (Harman 1987: 55, 58.)

2b. The contents of concepts are determined by their functional role in a person's psychology. (Harman 1987: 55.)

Now, the picture we get from (1) and (2) is shown in Figure 4.4. But this seems to multiply entities. Now, I have not been known to be bothered by such multiplications in the past. However, I fail to see what "content" contributes here, perhaps because I fail to see what it *is*. Nor do I understand what it means for content (whatever it is) to "determine" meaning. In fact, earlier formulations of Harman's theory were more streamlined:

> The relevant thoughts are to be identified, not in terms of truth conditions, but rather in terms of their potential role in a speaker's "conceptual scheme" .... The meaning of a sentence is determined by the role in a conceptual scheme of the thoughts that the sentence would normally be used to express. (Harman 1974: 10–11.)

My view is this:

R1. The meanings of linguistic expressions *are* the thoughts they express (so "meaning" and "expression" are inverses of each other).

R2. The content of a thought *is* its functional role.

The SNePS/Cassie picture is this:

S1. Cassie's understanding of a linguistic expression is the set of nodes she maps it into (the set of nodes she uses to model the expression).

S2. Those nodes play a functional role in her entire semantic-network mind.

Presumably, Harman's "concepts" are SNePS/Cassie's base nodes ("concepts are treated as symbols in a 'language of thought'" (Harman 1987: 56)), and Harman's "thoughts" are SNePS/Cassie's molecular nodes. This appears to be consistent with (2a), but (2a) is ambiguous: What is it that is constructed out of concepts: Is it *thoughts*? Or is it *contents* of thoughts? On my view, "thoughts" would be constructed out of (or, would be structured by) "concepts" as well as other "thoughts" (for example, Cassie's thought that Oscar believes that Lucy is rich is constructed out of the thought that Lucy is rich and concepts of Oscar, Lucy, and being rich). And, in contrast to (2b), the "meaning" (in one sense) of thoughts *as well as* of concepts is a function of their location in the entire network of thoughts and concepts.

There is, as I mentioned, a third part to Harman's theory:

3. Functional role is conceived nonsolipsistically as involving relations to things in the world, including things in the past and future. (Harman 1987: 55; cf. Harman 1982: 247; Harman 1988.)

Now, on the SNePS/Cassie, first-person, internal view, there may indeed be other aspects to the notion of the functional (or conceptual, or inferential) role of a concept or thought. There is, for instance, their role in action (cf. Kumar 1993abc, 1994ab; Kumar & Shapiro 1993), although this role might not be (or contribute) anything over and above the concept's location in the network (and might, in fact, depend entirely upon it). But I part company with Harman on point (3). *Nonsolipsistic* functional role is not something the agent can have access to. Point (3) takes a third-person viewpoint, not a first-person one. I am solely interested in what linguistic expressions mean *to the agent*, not what a third person says that they mean for the agent.

There remains, nonetheless, the question of the relevance of Sellars's "entry" and "departure" rules, which seem clearly to be links with the external world. They are part and parcel of another issue that Harman raises: the role of language in *thought* as opposed to *communication*. I do not deny that there are "connections between concepts and the external world" (Harman

1987: 80). I merely deny that such connections tell the *cognitive agent* anything about his or her language or concepts. *Maybe* such connections *do* tell a third person something, but they give no first-person information. (The 'maybe' has to do with the point made in §§2.7.1 and 2.8.2 that, at least, the third person is making connections between his or her *own internal representations* (a) of the other agent's concepts and (b) of his or her own internal model of the world.)

Curiously, the only connections Harman explicitly mentions are those between concepts and *words* and those between concepts and "normal contexts of functioning" (Harman 1987: 80 **[???] check Lepore, New Directions in SEmantics**). But the link to words is of only *causal* interest. From the SNePS/Cassie point of view, what's important is the *presence in* the internal semantic network of a `LEX` node; *how* it got there is irrelevant. (That's what methodological solipsism is all about; cf. Ch. 6.) Ditto for normal contexts of functioning: They may give the third person some information, but they avail the first person nothing.

Clearly, it's in the case of "communication" that these issues come to the fore, not the case of "thinking". Harman distinguishes these two uses of language, and finds the latter to be more basic. I agree (to a point), but why then does he care about the external links? Let's look a bit more closely.

The view of language as serving a communicative function *sounds* similar to David Lewis's notion of "language" as

> A social phenomenon which is part of the natural history of human beings; a speaker **[???]—check article; see ms for full reference and add to biblio** of human action, wherein people utter strings of vocal sounds, or inscribe strings of marks, and wherein people respond by thought or action to the sounds or marks which they observe to have been so produced. (Lewis 1975: 3 **[see Harman 144 for ref]**).

But Harman seems to mean something more restrictive, for there can be communication via a syntactic system that is *not* language—for example, Morse code (Harman 1987: 57).

What about the role of language in thought? Harman cites Chomsky (who in turn paraphrases Humboldt):

> [...] to have a language is to have a system of concepts[10]
>
>> and it is the place of a concept within this system (which may differ somewhat from speaker to speaker) that, in part, determines the way in which the hearer understands a linguistic expression ... [T]he concepts so formed are systematically interrelated in an "inner totality," with varying interconnections and structural relations ... [cf. a semantic network.] This inner totality, formed by the use of language in thought, conception, and expression of feeling, functions as a conceptual world [cf. Dennett's "notional world" (1982)] interposed through the constant activity of the mind between itself and the actual objects, and it is within this system that a word obtains its value ....
>
> (Harman 1975: 273; unbracketed ellipses in Harman's text.)

[10]These could be the *meanings* in Lewis's theory of "a language"; Lewis 1975: 000 **[???]**.

Elsewhere, he calls this use of language "calculation, as in adding a column of figures" (Harman 1982: 242; 1987: 56), commenting that conceptual-role semantics "may be seen as a version of the theory that meaning is use, where the basic use of symbols is taken to be in calculation, not in communication, and where concepts are treated as symbols in a 'language of thought'" (Harman 1982: 243). This is clearly a syntactic enterprise.

There is some unclarity, however, when Harman speaks of these two uses of "language" or of "symbols" (e.g., Harman 1987: 56). When he talks of "symbols", is he talking about external linguistic expressions? Or is he talking about the internal symbols of a language of thought? For SNePS, the nodes are symbols of a language of thought, and they represent propositions, thoughts, and concepts (cf. Shapiro & Rapaport 1991, Shapiro 1993). They can be used in "calculation" (for example, inference) *as well as* in communication (for example, language is generated from them, and they are produced from language). Linguistic expressions are also used in communication. In fact, they are the *vehicles* of communication. *What gets communicated*—what is carried by the vehicle—are thoughts and concepts (that which is represented by the nodes). But linguistic expressions are not normally used in internal calculation (though, of course, they *can* be, as when Cassie wonders what Oscar meant when he *said* 'all mimsy were the borogoves').

My view is that both "thinking" (or "calculating") and "communication" are equally important components. There are spoken and written expressions. And in Cassie's mind, there are mental concepts in correspondence with them. There are also speakers and hearers, each of whom communicates with others, and each of whom understands the other by means of a semantic interpretation of the other's spoken or written expressions in terms of their own concepts. And, *pace* Harman, thinking *is* communicating with oneself (cf. Harman 1982: 243): This is Cho's point (§2.7.1), and it works (in part) by the mechanism of "internal reference" discussed in §§2.7.1, 2.8.2, and 8.3.1.

Harman and I are, however, not so far apart: "a language, properly so called, is a symbol system that is used both for communication and thought. If one cannot think in a language, one has not yet mastered it" (Harman 1987: 57). So far, so good. But: "A symbol system used only for communication, like Morse code, is not a language" (Harman 1987: 57). What, then, about Searle-in-the-room's use of Chinese, for communication only; is that not the use of a language? The answer depends on how much of the story Searle told us. As I claimed in "Syntactic Semantics" and §1.2.4, above, he didn't tell us enough. Here's how I see it: Unless the symbols are part of a large network, they have no (or very little) meaning—and, to that extent, maybe Searle has a point. But the more they *are* used for calculation/thinking, the more language-like they are. And, I claim (and I think Harman would agree), they *have* to be part of such a large network, otherwise they could not be used to communicate. They have meaning if and only if, and to the extent that, they're part of a large network. Searle, it seems to me, denies that being part of a large network suffices to provide meaning. What conceptual-role semantics says is that that's the only way to provide it:

> . . . there are two uses of symbols, in communication and speech acts and in calculation and thought. (Nonsolipsistic) conceptual role semantics takes the second use to be the basic one. The ultimate source of meaning or content is the functional role symbols play in thought. (Harman 1987: 79.)

4.4 OBJECTIONS.

There has been a large number of objections to conceptual-role semantics. Let's see how powerful they are.

4.4.1 General Objections.

4.4.1.1 Qualia.

Harman (1982: 250–252) points out that one objection to conceptual-role semantics arises from puzzles about qualia—the qualitative "feels" or private, subjective experiences associated with, for example, pains, visual perception, and so on. The existence of qualia suggests that there is something over and above functional (or conceptual) role that is important for what a concept is.

Since I'll have a lot to say about qualia in §7.6.3, let me put this line of objection on hold. I'll just mention that I find the defenses of functionalism against the qualia puzzles to be quite workable (e.g., Shoemaker **GIVE REF**). I'll provide my own defense—viewing qualia as "implementation side-effects", hence not in need of being accounted for by functionalism—in §7.6.3.

4.4.1.2 Speech-act theory.

Harman also raises some potential objections from speech-act theory (1982: 252–255). But this is not a problem for SNePS/Cassie, since all speech acts have an origination in nodes, hence they do have a conceptual role to play.

Related to this is Harman's discussion of Grice (Harman 1987: 56–57). There are, at least, three distinct kinds of "meaning": (1) *natural* meaning (as in: smoke means fire; these are relations between elements entirely within the semantic domain), (2) *non*-natural meaning (as in: 'Feuer' means fire; this seems to be referential meaning, or "expression meaning"), and (3) non-natural *speaker* meaning ("what a speaker ... of certain symbols means"; but note that, on my theory—and possibly that of Bruner 1983 (see §5.3, below)—the speaker could mean one of his or her *concepts or thoughts* rather than something in the world). According to Harman, Grice claims that expression meaning can be analyzed in terms of speaker meaning. This seems reasonable. And, according to Harman, Grice further claims that speaker meaning can be analyzed in terms of the speaker's intentions to communicate. (I'll have a lot more to say about this in §9.4, when we look at the question of whether non-humans, such as apes (and computers), can use language.)

But, according to Harman, this last claim

> overlook[s] the meaningful use of symbols in calculation. You might invent a special notation in order to work out a certain sort of problem. It would be quite proper to say that by a given symbol you meant so-and-so, even though you have no intentions to use these symbols in any sort of communication. (Harman 1987: 57.)

But you *might* and *could* so use them. So, speaker meaning could, perhaps, be analyzed in terms of the *potential* for communication. Again, *pace* Harman (1987: 56), there seems to be no good reason to deny that "calculation" or thought is internal communication.

Now, Harman has an interesting, but flawed, point to make:

> Suppose you use your special notation to work out a specific problem. You formulate the assumptions of the problem in your notation, do some calculating, and end up with a meaningful result in that notation. It would be correct to say of you that, when you write down a particular *assumption* in your notation, you meant such and such by what you wrote: but it would be incorrect to say of you that, when you wrote the *conclusion* you reached in your notation, you *meant* so and so by what you wrote. This seems connected with the fact that, in formulating the *assumption* as you did in your notation, you *intended* to express such and such an assumption; whereas, in writing down the *conclusion* you reached in your notation, your *intention* was *not* **[ROMAN?]—check Lepore volume** *to express such and such a conclusion* but rather *to reach whatever conclusion in your notation followed from earlier steps by the rules of your calculations.* (p. 57; my italics.)

Harman's point is this: You can't *intend* the *conclusion*, since you haven't reached it yet! Intending to express a thought involves a "translation" or "mapping" *from* the thought *to* the notation. After the calculation (which is purely syntactic), you "translate" or "map" *from* the notation *to* the thought; so it can't have been the case that you *intended* to express that thought. So, you didn't mean what you wrote when you wrote the conclusion-expressed-in-the-notation.

But that's quite odd. Consider the old saying that I don't know what I think until I read what I wrote. We use language to "calculate", to think. Indeed, I *don't* intend my conclusions *before* I say them—I say them and come to believe them *simultaneously*. But they mean what they mean in the same way that things I *do* intend to say mean what *they* mean.

Harman continues the previous quotation as follows:

> This suggests that you mean so and so in using certain symbols if and only if you use those symbols to express the thought that so and so, *with the intention of expressing such a thought.* (Harman 1987: 57; my italics.)

But that's not so. The whole point of symbols and "calculation" is that once I intend a symbol to mean so and so, then that's what it will always mean (for me), whether or not I intend it at any given time. That's what enables me to say that the conclusion-expressed-in-the-notation means so and so. It's what enables me to (inversely) "translate" or "map" from the symbols to meanings (and back again) freely, with or without intentions to communicate.

So: the italicized intention-clause of the right-hand side of the biconditional in the previous quotation has to be modified, perhaps as follows:

> Cognitive agent c means that so and so in using certain symbols if and only if
>
> 1. c uses those symbols to express the thought that so and so, and
> 2. c once (or initially) had the intention of expressing such a thought.

Or perhaps a compositional theory of intending will do the job. Surely, each of the *basic* symbols *in* a thought mean something for me if and only if I use them to express a *concept* with the intention of expressing that concept. Compositionally, a thought-symbol means something for me if and only if I *can* use it to express a thought. Here, no *intentions* to express that thought are needed.

4.4.2 Specific Objections.

4.4.2.1 The objection from the existence of a shared external world.

One of the major claims against a conceptual-role semantics is that it ignores the contribution of a truth-functional semantics, the contribution of *reference*, the *fact* that there exists a real world out there that is shared by interlocutors. What *is* the contribution of truth-functional semantics and reference, and what *are* the arguments that (1) they are needed and (2) there exists a shared external world? Let's look at (2) first.

Clearly, that there is an external world is a fundamental *assumption*. There are, to be sure, G. E. Moore's arguments for it (Moore 1939), but they amount to little more than a statement of faith or a claim that in fact we assume that the external world exists or that we behave *as if* it existed. That's consistent with my version of conceptual-role semantics. What is *reference*, after all? A cognitive agent (for example, Cassie, or me) uses a term t to refer to some entity e in its (her, or my) visual field or in its (her, or my) knowledge base.

The case where t refers to an entity in a knowledge base is purely internal (cf. Rapaport 1988, §3.4, on deixis). Cassie (or I) refers by t to the entity e that she (or I) thought of once before. Oscar (or you), hearing Cassie (or me) use t is prompted to think of e_O, which is the object Oscar (or you) believes to be equivalent to (or the counterpart of) the one Cassie (or I) is thinking of, as in Figures 4.5 and 4.6.[11] Whether or not there is an actual object, α, in the external world[12] that corresponds to Cassie's e and Oscar's e_O is irrelevant to explaining the semantics of t. If there is such an α, then there is a correspondence relation between e and α (and an external referential relation between t and α). But that relation is not accessible to *any* mind (except possibly God's, if one wishes to view the external world as (within) God's mind).

In the case where t refers to an entity in one's visual field, t still internally refers to an internal representation, e, this time causally produced (perhaps) by some actual object α. If α exists, then when Oscar hears Cassie use t, Oscar, with luck, will take Cassie to be talking about e_O, which is equivalent to (or a counterpart of) (Oscar's representation of) Cassie's e, as in Figures 4.7 and 4.8. Here, that (or whether) α exists is irrelevant to the *semantics* of t, and is not accessible by any (human) mind. If Cassie's and Oscar's communicative negotiations (see Ch. 5) are constrained by the "behavior" of e and e_O, then they might hypothesize the external existence of a noumenal object α, but each of them can only deal with their phenomenal e and e_O, respectively.

Taken together, the knowledge-base and visual-field cases explain why and how a third person can "assign [Cassie's] predicates satisfaction conditions" (Loar 1982: 274–275). It also takes care of any *argument* that truth and reference are *needed*. Truth and reference, we assume, are there, but inaccessible. Hence, they *couldn't* be *needed*.

The contribution of truth and reference is by way of an *attempt* (doomed to failure) to

[11] I owe the style of picture to Perlis 1994.

[12] In Rapaport 1976, 1978, 1981, 1985/1986, I called this a "Sein-correlate".

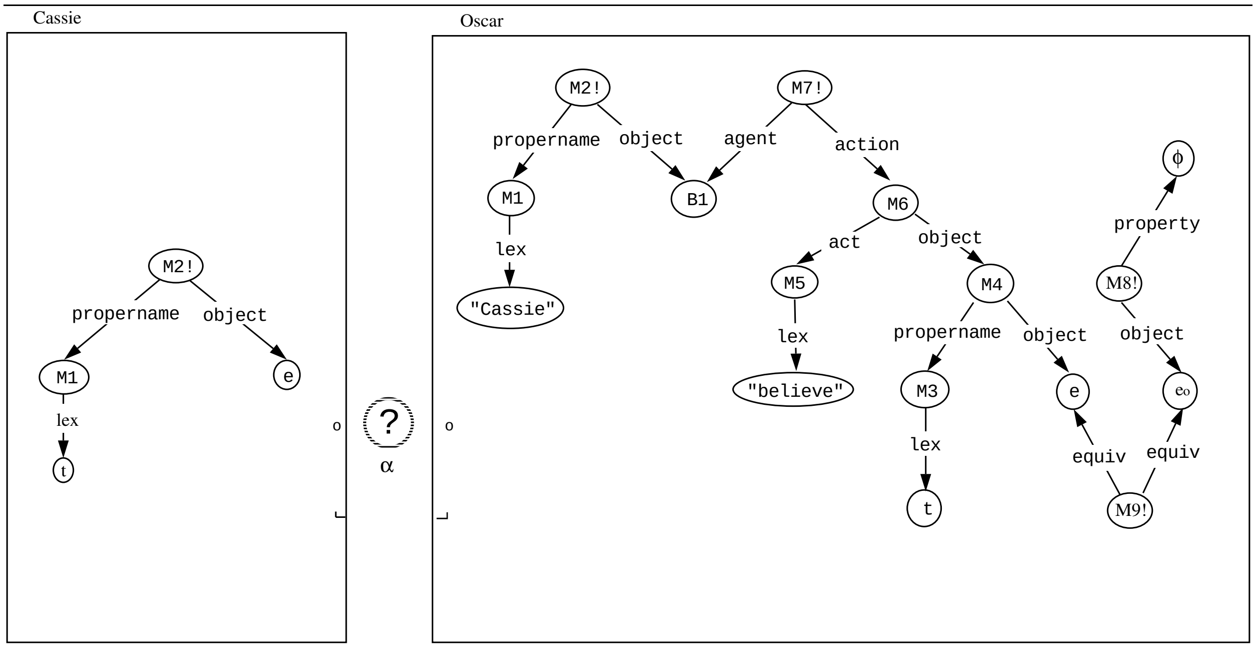


Figure 4.5: In Cassie's belief space: M2 = e is named 't'
In Oscar's belief space: M2 = B1 is named 'Cassie'
M7 = Cassie believes that M4
M4 = (Cassie's) e is named 't' (by her)
M8 = e_O has property φ
M9 = (Cassie's) e is equivalent to e_O
In the external world, α would be the object that Cassie thinks of as e (and refers to by t) as well as the object that Oscar thinks of as e_O, if it exists.

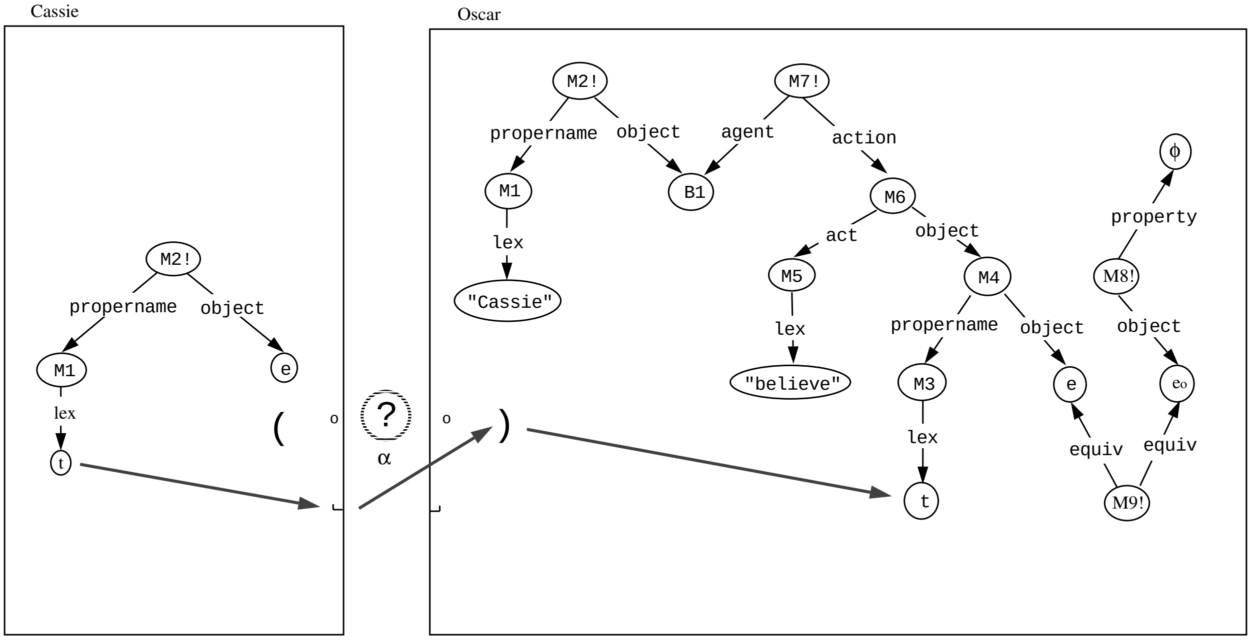


Figure 4.6: Detail of Figure 4.5. Cassie utters 't'; Oscar hears 't' and believes that Cassie is thinking of what Oscar thinks of as e_O.

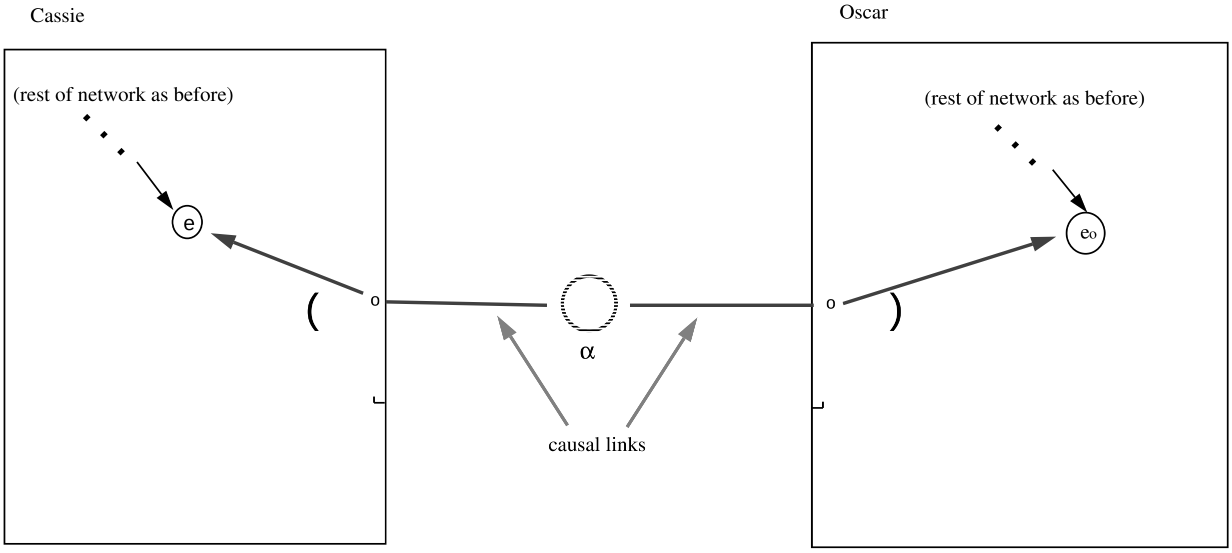


Figure 4.7: The visual-field case, with causal links.

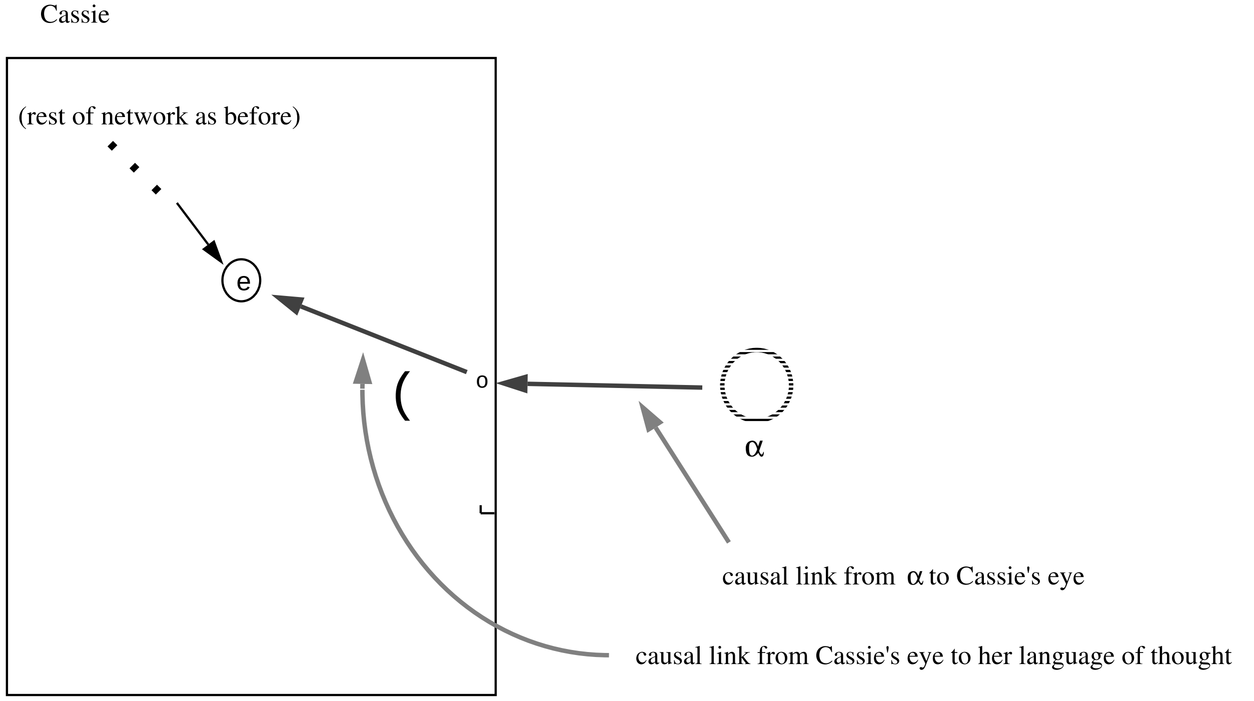


Figure 4.8: Detail of Figure 4.7. The causal link from α to e is here analyzed into two links, one from α to Cassie's eye, and another from her eye to her language of thought.

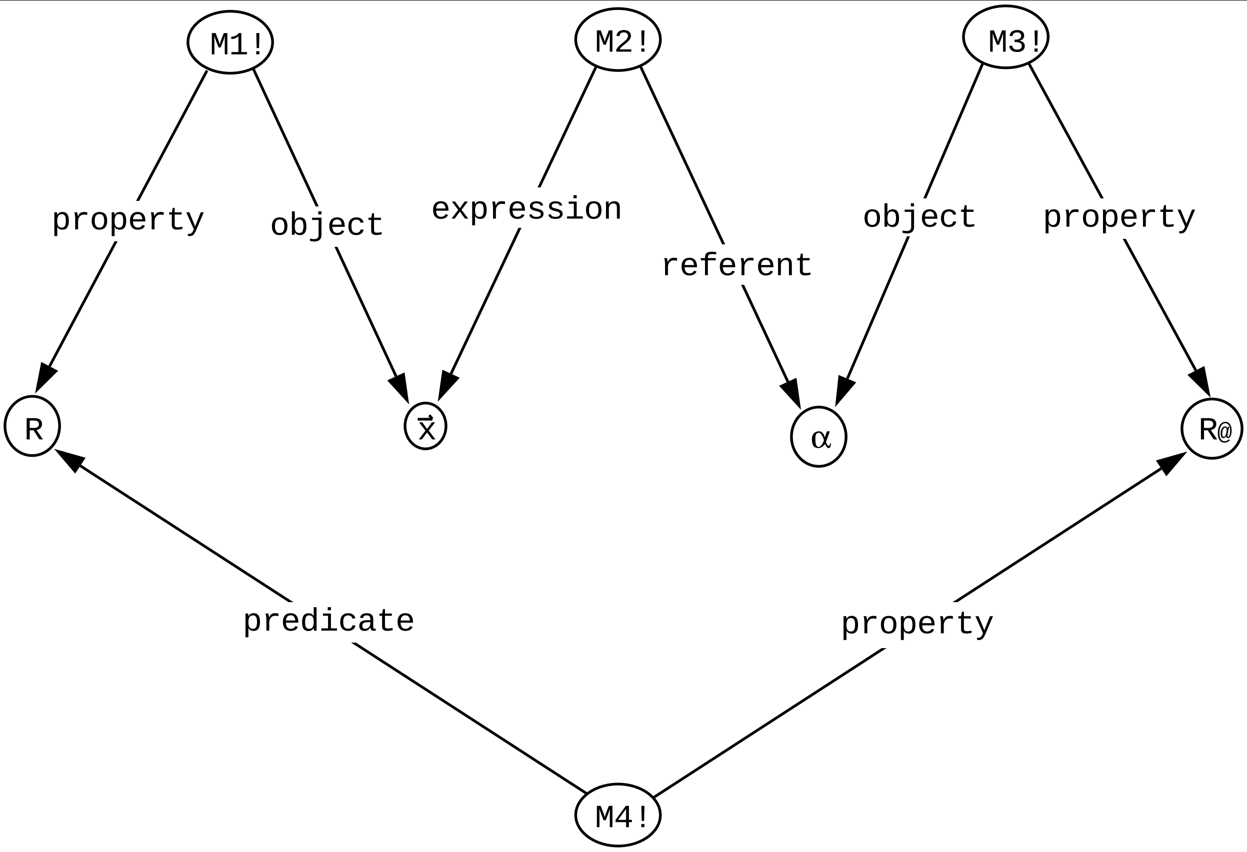


Figure 4.9: $\texttt{M1} = R\vec{x}$
$\texttt{M3} = R_{@}(\alpha)$
$\texttt{M2} =$ Expression $\vec{x}$ internally refers to α
$\texttt{M4} =$ Predicate R internally corresponds to property $R_{@}$

describe what the world is like. They are *metaphysical* notions. Recall that Cassie's claim that Oscar *knows* that φ is really just her claims that Oscar *believes* that φ and that she, too, believes that φ. Similarly, where α_i are "real-world" objects and $R_{@}$ is a "real-world" relation, her claim that $\ulcorner R(x_1, \ldots, x_n)\urcorner$ is true in the sense that $(\exists\alpha_1, \ldots, \alpha_n, R_{@})[R_{@}(\alpha_1, \ldots, \alpha_n)]$ is just her *belief* that $(\exists\alpha_1, \ldots, \alpha_n, R_{@})[R_{@}(\alpha_1, \ldots, \alpha_n)]$, as in Figure 4.9. That is, Cassie will have two "mental models": One is her mental model of the actual world; the other is her set of concepts about those things. Perhaps, as is my wont, I am multiplying entities. If so, that just strengthens my internalist perspective (for either R and x go, or $R_{@}$ and α go; what's left is still internal.

4.4.3 Lewis's Objections.

David Lewis's "General Semantics" (1972) is often cited in objections to conceptual-role semantics, or, more specifically, to theories of "semantic interpretation as the assignment to sentences and their constituents of compounds of 'semantic markers' or the like" (p. 169):

> Semantic markers are *symbols*: items in the vocabulary of an artificial language we may call *Semantic Markerese*. Semantic interpretation by means of them amounts merely to a translation algorithm from the object language to the auxiliary language Markerese. But we can know the Markerese translation of an English sentence without knowing the first thing about the meaning of the English sentence: namely, the conditions under which it would be true. (p. 169.)

But such a translation algorithm is *all* that Cassie (or any of us) *can* do. For Lewis, however, semantics consists of truth conditions. But how can Cassie come to know those without direct access to the external world? Perhaps she doesn't need such access. After all, she doesn't need to know the truth *value* of a sentence, only its truth *conditions*. But that, as we've seen, can be handled completely internally. How would Lewis distinguish *that* from Markerese?

Using Markerese is purely syntactic (pp. 169–170). So, ultimately, says Lewis, we need "to do real semantics at least for the one language Markerese" (p. 169). But how? Perhaps via names plus compositionality? If so, then except for the one-time causal production of an internal name by an external object, all is internal and syntactic. And why would we need "to do real semantics"? Perhaps to ground our internal symbols. But that can be done internally (as we saw in §5.2.2.2). **[???]**

Lewis makes much ado about the finitude of Markerese, which "prevents Markerese semantics from dealing with the relations between symbols and the world of non-symbols" (p. 170). Of course, as Smith has reminded us (see §2.7.1), semantics in fact does *not* deal with that relation or with "the world of non-symbols". Lewis's point is that "meanings may turn out to be ... infinite entities" (p. 170); our minds, however, are finite (cf. Smith's notion of "partiality", discussed in §2.2.2). The infinite entities that Lewis takes meanings to be are (roughly) intensions in the Montagovian sense: functions from indices to extensions (cf. p. 176). Presumably, since these take infinite possible worlds among the indices, they are infinite, hence could not be Markerese. But Markerese symbols could be finite specifications (indeed, algorithms) of such functions, for example, a propositional node (for example, `M7` in Figure 2.14) plus its surrounding network, together with the ATN parsing-generating algorithm, which "tells" Cassie how—or provides for her a method—to determine the truth *conditions* of 'Lucy pets a dog'.

'Truth' conditions are, however, a misnomer. Better to call them '*belief*' conditions: Cassie should believe 'Lucy pets a dog' if and only if she believes that `B1` represents an entity named 'Lucy', and she believes[13] that `B2` represents a member of the class of dogs, and she believes[14] that `B1` performs the actions of petting `B2`.

4.4.4 Potts's Objections.

Timothy Potts's essay "Model Theory and Linguistics" (1973) is instructive, because he agrees with much of what I have had to say yet still locates meaning in the world.

He begins by observing that in model theory, one "translates" one formal system to another "whose properties are already known .... [T]he systems thus related to the one under investigation are termed 'models' of it and known properties of the models can then be extrapolated to the new

[13] I.e., believes *de re*; that is, she need not have any beliefs about class membership.
[14] Again, *de re*; she need not have any beliefs about acts, actions, or their objects as such.

system" (p. 241). This is a clear statement of semantic understanding by general correspondence with an antecedently understood domain; anything, presumably, can be a model of anything else. The problem, as Potts sees it, is that model theory cannot provide a theory of *meaning* for natural language considered as a formal system. His argument is that (1) a theory of *meaning* requires a relation between a language *and the world*, not between two *languages*, and (2) model theory only gives a relation between two languages. Consistent with my support for conceptual-role semantics, I can accept (2), though I will reject (1). More interestingly, we will see that Potts's argument for (1) self-destructs! (Another argument Potts has is that natural language isn't a formal system in the first place. But it *is* a syntactic system, and that's all that's needed for the cases I am concerned with.)

1. First, some preliminary remarks to remind you of the theory I have been adumbrating. Aren't language-translation manuals theories of meaning of the one language in terms of the other? Recall from §3.2.2.2 that a speaker of English would be satisfied if told that the French word '*chat*' means "cat", while a speaker of French would be satisfied[15] if told that it means "*petit animal domestique, dont il existe aussi plusieurs espèces sauvages*". But 'cat' itself needs to be grounded in a definition of the form "that animal over there". But then we simply have a correspondence continuum: '*chat*' means (or is "grounded" in) 'cat', which in turn is grounded in that animal over there. To learn "the meaning" of '*chat*', one only has to stop at the first antecedently understood domain. And, in any case, "that animal over there" is at best an internal concept. The only "hooks onto the world" (Potts 1973: 241) are really hooks onto other *internal* nodes. So "that animal over there" is *really* a pointer—not to the world—but to an internal (non-linguistic) representation of the world, as we discussed in §§2.7.1 and 2.8.2 (cf. §8.3.1, below, and Perlis 1991, 1994).

2. Potts has some useful things to say about models. He first distinguishes

> between *being* a structure and *having* a structure: Something *is* a structure if it has distinguishable parts or *elements* which are inter-*related* in a determinate way. ... [T]wo different things, each of which **is** a structure, can in certain circumstances be said to **have** the *same* structure .... (p. 244; Potts's italics, my boldface.)

'Structure' seems intended as a neutral term; it is, in my terminology, a syntactic notion, since it refers to a system with "elements" that are "interrelated". To clarify this distinction, Potts discusses the example of a three-dimensional, cardboard model of a house and a two-dimensional blueprint as a model of a (possibly the *same*) house:

> Both the drawings and the cardboard model would then qualify as models of the building, each of them having a structure which is also a structure of the building. But now suppose that we have only the drawings and the cardboard model: the building has not yet been constructed. How can we say that they are models of a building, when there is no building of which they are models? and how can we say that they are models of the *same* building? ...
>
> These considerations show that the expression *is a model of* is, in logician's parlance, 'intensional'. Accordingly, we cannot say that what makes something which is a structure a model is that there is something else which is also a structure and that both have a structure in common. (p. 245.)

[15]Though I have my doubts!

That is, 'is a model of' is inten*s*ional in the sense that its second argument need not exist in the external world (cf., e.g., Rapaport 1985/1986). More to the point, however, is the fact that 'is a model of' is asymmetric. In any case, as I have argued elsewhere, the common *structure* can be taken as an inten*t*ional object (Rapaport 1978), and both the cardboard structure and the blueprint can be taken as models (actually, "implementations") of *it*. Nor does it *follow* from the intensionality of 'is a model of' that the cardboard structure is not a model of the blueprint. Clearly, it *can* be one, as long as the appropriate mappings (correspondences) exist (or can be defined).

3. Potts provides an argument for a conclusion that is close to Smith's gap between model and world. Potts's gap is between the language used to describe the model and the model itself:

> In [mathematical] model theory, the structures which are correlated with formal systems are **abstract** structures and thus inaccessible to perception. This is supposed to make no essential difference .... (p. 247.)

The situation with abstract structures, according to Potts, is that the abstract structure that is the model of the formal system is *not* directly correlated with it. Rather, the only way to access the abstract structure is via an antecedently understood meta-language for it, and it is the correlations between that meta-language and the formal system's object language that does the work:

> the abstract structure is a mere beetle in a box. ... We are not really studying the relations between a formal language and an abstract structure, but between two languages. Model theory is, rather, an exercise in *translation*. We have given meanings to the formulae of our object-language by specifying how they are to be translated into propositions of an **established** language with which it is assumed that we are **already familiar**; to this extent it is true that model theory is concerned with meaning. (p. 248; Potts's italics, my boldface.)

So, Potts has now argued for (2): model theory only gives a relation between two languages. I agree. He still needs to argue for (1): that even though such interlinguistic translation "is concerned with meaning" to some "extent", a *real* theory of meaning requires a relation between language and the world, that is, that meaning is *reference*, not sense or conceptual role. As I see it, of course, it's *primarily* sense or conceptual role. Why do I see it thus? For *de dicto*/intensional reasons: I'm concerned with the *beliefs of a cognitive agent*, not with whether those beliefs are true. Reference enters in two ways. (a) I explicate sense as a kind of reference to a domain of intensional entities (cf. §2.8.1, above). (b) Symbol grounding also requires a kind of reference, but this is a relation between internal nodes, only some of which are perceptually caused (§3.2.2.2.2).

4. Potts's argument for his claim that model theory doesn't do the job undercuts his claim about (1), for reasons not unrelated to Smith's gap:

> Thus it is just a confusion to suppose that model theory can say anything about the relation of language to the world; it can, at best, only elucidate one language by reference to another. This is all that is needed for its proper, mathematical application, for if the metalanguage is itself a formal language whose properties have already been studied, then the possibility of specifying a translation from the object to the metalanguage

> allows us to conclude that the object-language has corresponding properties. Talking of a structure in this connection is then quite harmless, though redundant. ... so the question whether ... expressions [of the meta-language] have a meaning by denoting [elements of the abstract structure] ... need not concern us. (pp. 248–249.)

This is astounding! For it can be taken to argue for our purely internal, methodologically solipsistic view by making three substitutions: (i) 'real world' for 'abstract structure' (after all, the real world is supposed to provide the semantic grounding for our language, just as a model is), (ii) 'Cassie's language' for 'meta-language', and (iii) 'Oscar's language' for 'object language'. That is, think of two cognitive agents, Cassie and Oscar, trying to talk about the shared external world by communicating with each other:

> [We] can, at best, only elucidate [someone else's] language by reference to [our own]. This is all that is needed for [understanding], for if [Cassie's language] is itself a formal language whose properties have already been studied[—that is, is antecedently understood, *syntactically*—]then the possibility of specifying a translation for [Oscar's language] to [Cassie's language] allows us to conclude that [Oscar understands things as Cassie does]. Talking of [the real world] in this connection is then quite harmless, though redundant. So the question whether [Cassie's language has] a meaning by denoting [things in the real world] need not concern us.

Syntax plus successful communication suffices for semantics. (We'll return to this theme in §5.3.)

4.4.5 Loewer's Objections.

Barry Loewer's essay, "The Role of 'Conceptual Role Semantics'" (1982, cf. Lepore & Loewer 1981), offers a Davidsonian argument that truth-conditional semantics "will provide the core of an account of the understanding of language used in communication" (p. 307). Here is my reconstruction of his argument.[16] Consider the following reasoning to justify a conclusion that it's snowing:

[16] "Arabella, Barbarella, and Esa are in a room with Arabella looking out the window. Arabella and Barbarella understand German but Esa does not. Arabella turns from the window to Barbarella and Esa and utters the words 'Es schneit'. On the basis of this utterance Barbarella comes to believe that it's snowing (and also that Arabella believes that it's snowing, etc.) while Esa comes to believe only that Arabella said something which is probably true. We can focus on the question of what knowledge comprises Barbarella's understanding 'Es schneit' by asking what would Esa need to know to come to the same beliefs as Barbarella. The obvious candidate for this knowledge is the knowledge that 'Es schneit' is true iff it's snowing. ... A reconstruction of the reasoning which justifies Barbarella's acquisition of the belief that it's snowing looks like this:

1. Arabella utters the words 'Es schneit'
2. Since 'Es schneit' is an indicative sentence and since Arabella is generally reliable, her utterance of 'Es schneit' is true
3. 'Es schneit' is true iff it's snowing

therefore

4. It's snowing.

Both Esa and Barbarella can come to believe that 'Es schneit' is true by knowing a bit of German grammar (enough to recognize indicative sentences) and knowing that Arabella is reliable. But only Barbarella is in a position to go on to conclude that it's snowing since only she understands German. And if my argument is correct that [sic; should be 'then'?] understanding must consist in part in knowing the truth conditions of the German sentence" (Loewer 1982: 306–307).

1. Arabella, a German-speaker, looks out the window and utters "*Es schneit*".
2. (a) '*Es schneit*' is an indicative sentence.
   (b) Arabella is generally reliable.
   (c) $\therefore$ Arabella's utterance of '*Es schneit*' is true.
3. '*Es schneit*' is true if and only if it's snowing.
4. $\therefore$ It's snowing.

Now, (4) is supposed to be the conclusion that Arabella's German-speaking listener, Barbarella, comes to. Here, truth conditions appear to play an essential role in the inference to (4), that is, in Barbarella's understanding what Arabella said. In contrast, Arabella's non–German-speaking listener, Esa, does not conclude (4), presumably because he does not know the truth conditions. But let's consider Barbarella's and Esa's cases separately.

Case 1: Barbarella. What is it that Barbarella comes to believe after (1)? Answer: a belief that it is snowing, that is, a belief that she, too, would express as '*Es schneit*'. She believes the *proposition*, not the utterance (cf. Shapiro 1993); at least, let's suppose so, though in this case it doesn't matter.

But she doesn't have to arrive at that belief by believing (3). Take her first-person point of view: She hears '*Es schneit*'; she processes it as an indicative sentence, and she constructs a mental representation of the proposition it expresses. She believes that proposition because of (2b). Thus, neither (2c) nor (3) are needed!

Moreover, (2c) follows from (2a) and (2b) by some rule such as this:

(i) Indicative sentences uttered by generally reliable people are true.

That seems wrong: Generally reliable people can be mistaken. For instance, Arabella might, without realizing it, be looking at a movie set with fake snow; or Barbarella might not realize that Arabella is acting in the movie and merely uttering her lines! However,

(ii) Indicative sentences uttered by generally reliable people are believable (or: ought, *ceteris paribus*, to be believed).

seems more reasonable and all that is needed for Barbarella to come to believe that it is snowing. So (3) is not needed at all. And neither, then, is truth-conditional semantics needed to account for the communicative use of language (or, at least, Barbarella's communicative use).

Case 2: Esa. Loewer ignores Esa, except to say that all Esa comes to believe is that *what Arabella said* (whatever it meant) is probably true. On my view, Esa comes to believe not that but, rather, that he ought to believe what Arabella said (even though he doesn't know what that is). Once again, truth conditions are not needed.

But suppose that Esa, although not a native speaker of German (like Arabella and Barbarella), is *learning* German and can translate '*Es*, '*schneit*', and N+V sentences into, say, English. Then Esa can reason more or less as follows:

1. Arabella uttered '*Es schneit*' (as before).

2. (a) '*Es schneit*' is an indicative sentence.
   (b) Arabella is generally reliable.
   (c) ∴ Arabella's utterance ought to be believed (*ceteris paribus*).

3. '*Es schneit*' means (i.e., translates as) "It's snowing".

4. ∴ I ought to believe that it's snowing (*ceteris paribus*).

Step (3) should be understood, not as saying that the German expression '*Es schneit*' means the English expression 'It's snowing', but as saying that '*Es schneit*' means the same thing as 'It's snowing', where 'It's snowing' means (say) `M1`—where, finally, `M1` is a mental representation in Esa's language of thought. Again, there is no need for truth conditions.

Another possibility is that Esa doesn't speak German, but also looks out the window and (somehow) infers or makes an educated guess that '*Es schneit*' expresses the weather. Since Esa sees that it's snowing, he infers or makes an educated guess that '*Es schneit*' means that it's snowing. Again, there is no role for truth conditions to play *in accounting for communicative understanding.* More precisely, there is no role for *external* truth conditions (which is the sort that Davidson, Loewer, et al., are talking about). Arguably, Esa's internal representation of the fact that it's snowing plays the same role *internally* that external truth conditions would play in the Davidsonian/Loewerian story. But this is akin to internal reference. It is all internal, and all syntactic.

Let me conclude my discussion of Loewer with one more lengthy quotation with which I *almost* agree:

> The question of how one understands the language one thinks in does seem to be a peculiar one. ... CRS [conceptual-role semantics] clarifies the situation. It is plausible that understanding a certain concept involves being able to use that concept appropriately. For example, to understand the concept red is, in part, to be able to discriminate red things. According to CRS an expression in P's Mentalese has the content of the concept red just in case it plays the appropriate role in P's psychology, including his [sic] discriminating red things. It follows that if some expression of P's Mentalese is the concept red then P automatically understands it. The answer may appear to be a bit trivial—P understands the expression of his Mentalese since if he didn't it wouldn't be his Mentalese—but it is the correct answer. If there are any doubts compare the questions we have been considering with "In virtue of what does a computer 'understand' the language it computes in?" Of course the understanding involved in understanding Mentalese is different from the understanding one has of a public language. I argued that understanding the latter involves knowing truth conditions. Not only would knowledge of truth conditions contribute nothing to explaining how we understand Mentalese but, it is clear, we do not know the truth conditions of Mentalese sentences. (Or, for that matter, even the syntax of Mentalese.) If P were to encounter a sentence of Mentalese written on the wall (in contrast to its being in just the right place in his brain), he wouldn't have the vaguest idea of what it means because he does not know its truth conditions. (p. 310.)

There is much to agree with here—except, of course, that understanding a public language, as *I* have argued, does *not* "involve knowing truth conditions" (except in the sense, which Loewer would not accept, that Esa, above, might have "internal truth conditions"). P's "automatic" understanding of expressions of his Mentalese is just what I have been calling "getting used to", that is, syntactic understanding.

What about Loewer's last claim, that "If P were to encounter a sentence of Mentalese written on the wall ... he wouldn't have the vaguest idea of what it means because he does not know its truth conditions"? Consider Cassie. She, too, has no knowledge of her language of thought, no knowledge of nodes, arcs, or arc labels. Only if she were a cognitive scientist and had a *theory* of her understanding would she be able to go beyond mere syntax. Even so, it would all be internal: Her theory that her belief that, say, Lucy is rich had a certain structure of, say, nodes and labeled arcs would be expressed in her language of thought. She might, for example, believe (correctly) that her belief that Lucy is rich consisted of two propositions: that someone was named 'Lucy' and that that someone was rich. In turn, she might believe (correctly) that the first of these had the structure that an object had a proper name that was lexically expressed by 'Lucy' and that the second had the structure that that object had a property lexically expressed by 'rich'. But her belief that this was so would involve her having *nodes* corresponding to the *arcs* of her actual belief, as in Figure 4.10. It is all internal, and it is all syntactic. Unlike the case of the Earth, which rests on the back of an elephant, which in turn rests on the back of a turtle, after which it's turtles all the way down—unlike that, it's *not* syntax all the way down. It stops, normally at the level of Cassie's language of thought, which she understands syntactically—that is, which she uses. If she were a cognitive scientist, she might devise a theory of her language of thought, and show the correspondences (as in Figure 4.10). Could she have a theory of the theory of her language of thought? That is, could she talk about the labeled arcs used in that theory? Only by means of "modifying" them. But there will always be more arc labels about which she cannot talk (and of which, in good Wittgensteinian fashion, she must be silent). We have a final turtle—we stop a Bradleyan regress (or is it a Platonic third-man?)—by not trying to give a *semantical* understanding.

There are further complications. Cassie's theory of her language of thought need not be a theory *about* arcs and nodes. It might, heaven forbid, be a connectionist theory! Even if her theory *were* about arcs and nodes, and even if her theory of representation matched her actual representations (as opposed, say, to a representation using the theory of Richard Wyatt (1989, 1990, 1993), still she would not be able to supply "truth" conditions, since she would not be able to *mention* (but only *use*) her own representations. Only a third person—a computational neuroscientist—could determine whether her theory were true—that is, could determine whether the representations of her theory corresponded to her actual representations. (And then, of course, this could only be done internal to the computational neuroscientist's own mind—but I won't press that point here.)

4.4.6 Lycan's Objections.

William G. Lycan defends the need for truth conditions in his *Logical Form in Natural Language* (1984), arguing that truth plays a role in the translation from utterance to Mentalese:

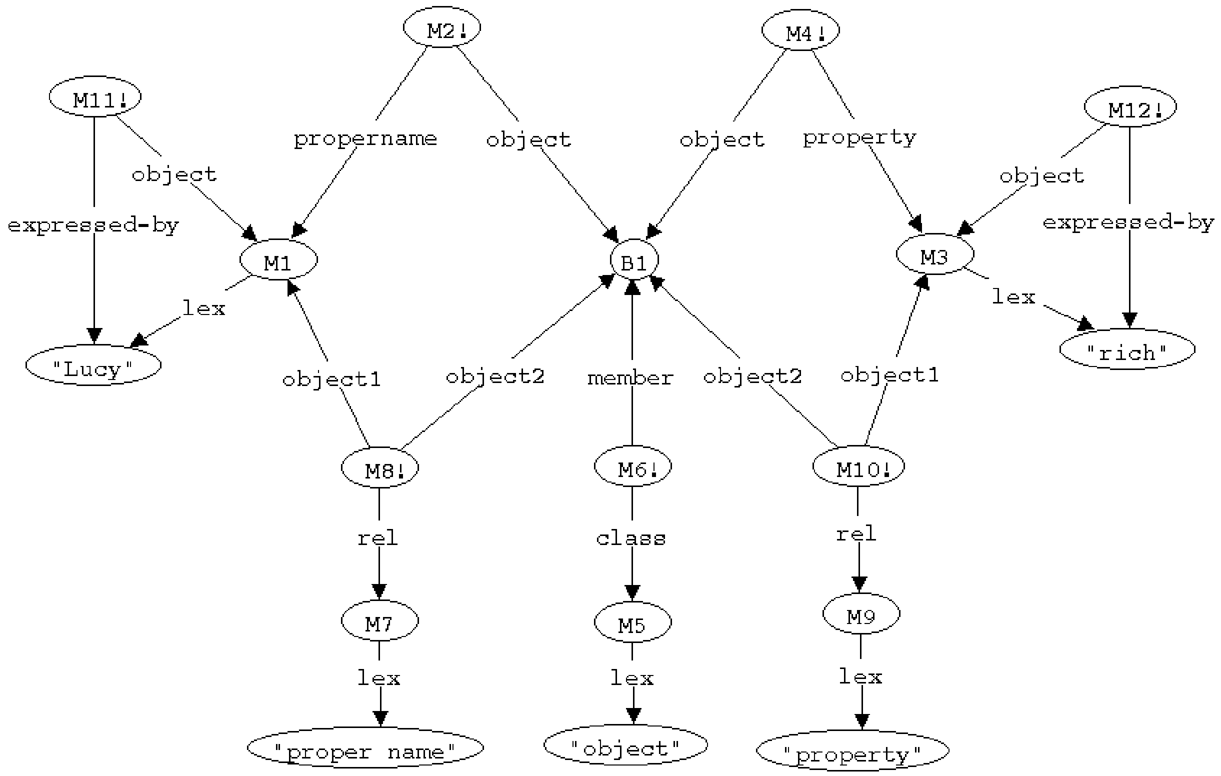


Figure 4.10: `M2!` = `B1` is named 'Lucy'
`M4!` = `B1` is rich
`M6!` = `B1` is a (member of the class of) objects
`M8!` = `B1` is related by the `propername` relation to `M1`
`M10!` = `B1` is related by the `property` relation to `M3`
`M12!` = `M3` is lexically expressed by 'rich'
`M11!` = `M1` is lexically expressed by 'Lucy'

> If a machine or a human hearer understands by translating, how does the translation proceed? Presumably a recursion is required .... And what property is the translation required to preserve? *Truth together with its syntactic determination* is the obvious candidate. Thus, even if one understands in virtue of translating, one translates in virtue of constructing a recursive truth theory for the target language. (p. 238.)

Now, the translation may *in fact* preserve truth. I don't deny that there is such a thing as truth (or reference), only that it's not needed to account for how we understand language. But the translation algorithm (the semantical procedure of procedural semantics) makes no more explicit *appeal* to truth (to truth *values*) than do rules of inference in logic. Truth can be used to *externally justify* or *certify* the algorithm (or the rule of inference), but the translation (or the inference) goes through anyway, in a purely syntactic fashion.

Negotiation, however, does play a role in adjusting the translation. In fact, the translation might *not* preserve truth. But the process of language understanding is self-correcting.

> ... the assignment of full-fledge[d?] **[???]** truth-conditions to sentences of a natural language helps to explain why a populations' having that language confers a selectional advantage over otherwise comparable populations that have none (this point is due to Dowty ...) .... (p. 240.)

I take this to be part of "negotiation"—only here it's negotiation with the *world*. Is it possible that (1) the need to accept the existence of others with whom we communicate and the existence of the world and (2) the need for negotiation just *is* the claim that truth-conditional semantics plays a role in our understanding of natural language? Sellars and Harman don't think so:[17] They allow for language-entry/exit rules. If (1) and (2) *do* amount to the need for truth-conditional semantics, then I suppose we're just differing on, excuse the expression, semantics, and I probably am taking an intermediary position à la Loewer et al. Still, from the first-person point of view, given that there *is* external input, the rest of the story is all internal. We'll return to the issue of negotiation in Chapter 5.

4.4.7 Fodor and Lepore's Objections.

In "Why Meaning (Probably) Isn't Conceptual Role" (1991),[18] Jerry Fodor and Ernest Lepore argue, not that conceptual-role semantics is *wrong*, but that it is inconsistent with two other principles that normally accompany it: compositionality and the analytic–synthetic distinction (p. 332). Now, personally, I like all three. So am I doomed to inconsistency? I'd like to think not. Let's see.

Fodor and Lepore begin with an assumption (which suggests that the inconsistent triad of conceptual-role semantics, compositionality, and the analytic–synthetic distinction may, rather, be an inconsistent *tetrad*) "that the fact that a word ... means what it does can't be a brute

[17] Or maybe they do—cf. Harman on wide functionalism and my reply to that, §6.5.2.

[18] In addition to the provenance of this paper as given in Fodor & Lepore 1991: 328fn (i.e., adapted from Fodor & Lepore 1992 (cf. their Ch. 6) and originally presented at the 1991 Chicago Linguistic Society (and published in its proceedings), it was also read by Fodor at the SUNY Buffalo Center for Cognitive Science Conference on Cognition and Representation (April 1992).

fact. ... [S]emantic properties must *supervene on* nonsemantic properties" (p. 329; for more on supervenience, see Chapter 7). But why? Or better: What does this mean? It doesn't mean "that semantic properties ... [are not] irreducibly intentional, or irreducibly epistemological, or irreducibly teleological" (p. 329). It does mean that "It can't be a brute fact ... that 'dog' means *dog* and not *proton* and that 'proton' means *proton* and not *dog*" (p. 329).

Why can't that be a brute fact? It's certainly an *arbitrary* fact; for example, 'dog' doesn't *resemble* dogs. So 'dog' *could* have meant "proton" or even "cat". Why *does* 'dog' mean "dog"? The story is, no doubt, buried in pre-etymological history, but one can guess that at some time, someone said 'dog' (or some etymologically-related ancestor) when in the presence of a dog. Now, why so? Isn't that a brute fact? And, if so, it certainly seems to be a *semantic* fact in just about every sense of that term, including that of correspondence. It is, no doubt, also an intentional (or perhaps epistemological or teleological) fact, but perhaps that's just what it is to *be* a semantic fact.

Now, as it happens, just this story is cited by Fodor and Lepore as an example of a *non*-semantic answer (p. 330). It's one of the versions of what they call "Old Testament" semantics, "according to which the meaning of an expression supervenes on *the expression's relation to things in the world*" (p. 329). Now, I certainly am not an Old Testament semanticist. That is, although I recognize that there was, at some time, a causal link between dogs and 'dog', no doubt mediated by an internal mental representation of a dog, nevertheless that's not, for me, the fundamental meaning of, say, *my* use of 'dog'. For one thing, I might never have seen a dog; I've certainly never seen an aardvark, or a proton, or a unicorn, yet the words for dogs, aardvarks, protons, *and* unicorns are equally *and in the same kind of way* meaningful to me. So their meanings must have to do with something other than (perceptual) experiences of them. But even if I *were* an Old Testament semanticist, I'd consider the dog–'dog' relation to be a semantic one, and brute at that. (For another thing, as Fodor and Lepore point out, there are the Fregean 'morning star'–'evening star' cases, where Old Testament semantics would count these as strictly synonymous, though clearly they are not.)

By contrast, there is "New Testament" semantics, that is, conceptual-role semantics (although Fodor and Lepore see it more as *inferential*-role semantics, albeit broadly construed; cf. pp. 330–331).[19] According to New Testament semantics, semantics supervenes on "intralinguistic relations" (p. 332). And, as I have been at pains to convince you, such relations are indeed semantic—they're the base case of a recursive conception of semantics. But if Fodor and Lepore want to consider this as *non*-semantic, I'm willing to grant them this assumption for the sake of argument, and go on to bigger things.

4.4.7.1 The alleged evils of holism.

One of the bigger things is that conceptual-role semantics entails holism, which Fodor and Lepore see as a bad thing (p. 331). I, however, rejoice in the entailment. Why is conceptual-role semantics holistic? Because if an expression's meaning is its conceptual or inferential role in the language, it must be its *entire* role in the *entire* language, not some (arbitrary) subpart of either, and that's what holism is. Why is it supposed to be bad? Because it follows

[19] It is interesting to note that the title of their paper uses 'conceptual', yet their arguments are really about *inferential*-role semantics. This was first pointed out to me by Toshio Morita.

> that no two people ever share a belief; that there is no such relation as translation; that no two people ever mean the same thing by what they say; that no two time slices of *the same* person ever mean the same thing by what they say; that no one can ever change his [sic] mind; that no statements, or beliefs, can ever be contradicted ...; and so forth. (p. 331.)

Perhaps these do follow; but why are they *bad*? Or, rather, can we find the silver lining in this dark cloud? Let's consider these one by one.

1. *No two people ever share a belief:* This is false.[20] There *is* a way for two people to share a belief. If Cassie believes that Lucy is rich, and if Oscar also believes that (the same) Lucy is rich (and if their languages of thought express these beliefs in the same way),[21] then they share that belief. The essential core of the belief, the way it is represented or expressed, its *intrinsic features*, are identifiable independently of its place in the network and is common to its "instantiations" in Cassie and Oscar. Now, of course, if Cassie, but not Oscar, believes, in addition, that Lucy is tall, or if Oscar, but not Cassie, believes, in addition, that rich people are snobs, then the (inferential) roles of their beliefs will differ, and, so, the meanings of their utterances that "Lucy is rich" will differ. That is, the *relational properties* of the two "instantiations" differ, so their roles differ. Hence, by conceptual-role semantics, their meanings differ. That's another matter (see (3) below). But Cassie and Oscar do share a belief.[22]

2. *There is no such relation as translation:* This is, indeed and alas, true, if by that is meant something like literal, word-for-word, expression-for-expression, yet idiomatic translation with no loss of even the slightest connotation. Languages are just too subtle and complex for that. Literary translation is an art, not a science (cf. e.g., *The New York Times Book Review* (26 April 1992)). **[???]** True, '*Es schneit*' or '*il neige*' seem to translate pretty well as 'it's snowing'. (Or do they? Would 'it snows' be better? Arguably not.) But how about '*Pierre a un coeur de pierre*'? "Peter has a heart of stone" misses the pun. The trouble is that the networks of associations for any two languages differ so much that the conceptual roles of its expressions must differ, too. So, translation is out; paraphrases or counterparts are the best we can get. But at least we can get those.

3. *No two people ever mean the same thing by what they say:* This is true. *Your* utterance of 'Lucy is rich' does *not* mean what mine does, because of the differing conceptual roles each plays in our network of concepts. Yet we do manage to communicate. How so? Recall, first, Bertrand Russell's observation that if we *did* mean exactly the same things by what we said, there would be no *need* to communicate (1918: 195–196). So *lack* of exact synonymy may be a necessary precondition for communication. If you tell me "Lucy is rich", I understand you by mapping your utterance into my concepts. Since we speak the same language and live in the same culture, we share a lot of the same concepts, so the mapping is usually pretty good, though never perfect. Witness Cassie and Oscar above: For Cassie, a tall person is rich

[20] Or perhaps ambiguous, depending on what 'share' and 'belief' mean (and possibly on how one feels about connectionism.

[21] Their languages of thought may differ, of course, but I take it that that's not the point Fodor and Lepore are making.

[22] Or else case (1) is like cases (2) or (3).

(but not necessarily a snob); for Oscar, Lucy is a snob (but not necessarily tall). Though we understand *slightly* different things by what we each say, we understand nonetheless.

Suppose, however, that we don't understand each other. Suppose I think that 'punt' means "kick the ball and see where it lands" (or suppose that I have no idea *what* it means other than in the football *metaphor* "we'll just have to punt", uttered, usually, in circumstances where we're going to try to do something and if it fails, "we'll just have to punt" (that is, we'll have to figure out what to do at that time). (Clearly, I *don't* understand what it means!) Now suppose that I say "if this plan fails, we'll just have to punt" to you, but you *do* understand what it means and take me to be telling you that if what we try fails, then *you'll* have to find a solution. Clearly, we've failed to communicate if that's not what I intended. Equally clearly, a bit more discussion on our parts can clarify the situation, can help each of us readjust our networks: "Oh, what *you* meant by 'punt' is X"; "Oh, what *you* meant by 'punt' is Y, and you know better than I do, since I don't follow football, so, from now on, that's what *I*'ll mean by 'punt', too". This permits us to understand each other, *even though we don't ever mean (exactly) the same thing by what we say.*

4. *No two time slices of the same person ever mean the same thing by what they say:* This is also true, *mutatis mutandis*. In this very sentence that you are now reading, I don't mean by 'mean' what I meant in the previous sentence, since that was uttered by an earlier time slice of me, who didn't have *this* sentence as part of his network. Indeed, the previous sentence *extends* the conceptual-role–semantics meaning of 'mean'. Nevertheless, there's enough of an overlap for communication to succeed. Since this is the first-person case, however, and I'm mostly interested in the first-person case, let's consider it a bit further.

   One way to clarify the problem is to explicate the conceptual role of an expression E as the *set* of "contexts" it "appears in". For a concrete instance, in the SNePS case, this could be the set CR_E of all nodes that dominate or are dominated by the node for the concept expressed by E. (That set may well turn out to be the entire network, not necessarily excluding the nodes for the concept and expression themselves.) Now, suppose Cassie hears a new sentence that uses E. Then E's conceptual role changes to a *new* set, $CR'_E = CR_E \cup S$, where S is the set of all the nodes newly dominated by and dominating the E-node. Since sets are extensional beasts, $CR_E \neq CR'_E$. This, I take it, is the problem that Fodor and Lepore see.

   I think there are two ways out of it. One I sketched some time ago in "How to Make the World Fit Our Language" (Rapaport 1981): As the conceptual role of an expression grows, some parts of it will be seen as more central and, indeed, more stable than others. (Cf. Quine's "web of belief" (1951, §6); Ehrlich & Rapaport 1992, 1993, 1995; Ehrlich 1995.) Such a central, stable, dictionary-like "definition" of an expression will serve to anchor both int*er*personal communication and int*ra*personal meditation. After all, we don't normally bring to bear *everything* we know about a concept when we hear, use, or think about it.

   The other way out involves using the techniques of non–well-founded set theory to provide a stable identification procedure for nodes in ever-changing (or even circular) networks (Hill 1994).

5. *No one can ever change their mind:* This is false. As (4) shows, it's far from the case that no one can change their mind. Rather, everyone *always* changes their mind (literally, in the case of Cassie). But *that's* not a problem, for the reasons given in (4). (And, anyway, how does this follow from holism?)

6. *No statements or beliefs can ever be contradicted:* This is certainly also false: After all, we reason non-monotonically and are always, as noted in (5), changing our minds (Martins & Shapiro 1988, Martins & Cravo 1991). (And, anyway, how does *this* follow from holism?)

4.4.7.2 Compositionality and the analytic–synthetic distinction.

So, there's no reason to reject conceptual-role semantics just because it entails the alleged evils of holism. Is there, then, as Fodor and Lepore want to argue, reason to reject it on the grounds of inconsistency with the hypotheses "that natural languages are compositional, and ... that the a/s [analytic–synthetic] distinction is unprincipled" (in the sense "that there aren't any expressions that are true or false solely in virtue of what they mean") (p. 332)?

A preliminary remark before we look at Fodor and Lepore's argument. For me, truth and falsity are irrelevant, of course. So perhaps I have an easy way out: Give up the analytic–synthetic distinction on the grounds of irrelevance. But I suspect that there's a *doxastic* way to view the analytic–synthetic distinction that can avoid the need to deal with truth values yet still be, potentially, inconsistent with conceptual-role semantics and compositionality: Are there expressions that ought to be believed solely in virtue of what they mean? I suspect that the class of such expressions would be identical to the class of analytic expressions as Fodor and Lepore would characterize them. Thus, if 'bachelors are unmarried' is supposed to be true by virtue of the meanings of 'bachelor' and 'unmarried' (and 'are'), then and only then ought it to be believed for that reason. (For the record, I think it's not analytic either way you look at it.) Likewise, if one ought to believe 'red squares are red' solely in virtue of the meanings of 'red' and 'square' (and 'are'), then and only then is it true in virtue of those meanings. (And, for the record, I think this *is* analytic.) In what follows, then, I'll treat the analytic–synthetic distinction doxastically.

4.4.7.2.1 ***Compositionality.*** Consider, first, conceptual-role semantics and *compositionality*. Fodor and Lepore take compositionality to be "non-negotiable", since it is the only hypothesis that entails "productivity, systematicity and isomorphism", all of which they take as essential features of natural language (pp. 332–334). Compositionality, of course, only holds for non-idiomatic expressions, as Fodor and Lepore note. To say that, however, is to come dangerously close to circularity. For to say that compositionality only holds for non-idiomatic expressions is to say that it only holds for expressions that can be analyzed, that is, expressions whose meaning *is* determined by the meanings of its parts. So, compositionality only holds for expressions for which it holds. Having said this, however, I should also say that it certainly seems to be a reasonable principle, though I can easily imagine that a sustained effort to understand the semantics of idioms and metaphors (broadly construed after the fashion of Lakoff 1987) might undermine it. However, it hasn't, yet.[23] **[???]**

Productivity certainly seems to be a fact about languages, even *non*-natural ones. A non-compositional language would appear to need an infinite set of primitive terms or an infinite set of formation rules to be productive, and natural languages are clearly finite in both these respects, so finite, non-compositional languages would not be productive.

Systematicity, too, seems a general feature of languages and to follow from compositionality: If the meaning of, say, 'aRb' were *not* a function of the meanings of 'a', 'R', 'b', and of its formation

[23] But cf. Pelletier 19xx for arguments *against* compositionality.

rule, then there would be no reason to expect '*bRa*' to be well formed or meaningful (though it *might* be).

Isomorphism, however, *seems* a bit more suspect (as even Fodor and Lepore admit, p. 333n2). For one thing, Fodor and Lepore express it in a curiously, albeit apparently harmlessly, one-sided way:

(I) If a sentence S expresses the proposition that P, then syntactic constituents of S express the constituents of P. (p. 333.)

What about *vice versa*? Well, if a proposition, P, *has* constituents, and if each of them is expressed by (sub-sentential) symbols, then—by compositionality—it does appear that a sentence S so structured expresses P. But does P have to have constituents? What if *propositions* were unanalyzable units? Then the *converse* of (I) would be vacuous, I suppose. But that would play havoc with (I), itself: For S might *have* constituents, yet they could not, then, express P's constituents, since P wouldn't have any. Here's where compositionality comes to the rescue, I suspect. What is a proposition, anyway, and what does it have to do with compositionality? Well, compositionality as Fodor and Lepore have it says that *the meaning of a sentence* is a function of its syntactic structural description together with *the meanings of its lexical constituents* (p. 332). The link to propositions must be this: The meaning of a sentence is the proposition it expresses. In that case, lexical meanings must be constituents of propositions. So, compositionality entails that propositions are analyzable. I was willing to grant them that anyway, but I thought it was worthwhile to spell things out.

Here's the first problem (p. 334):

1. Meanings are compositional.

2. Inferential roles are not compositional.

3. ∴ Meanings can't be inferential roles.

We've just accepted (1). Must we accept (2)? Here's the first part of Fodor and Lepore's defense of (2): By compositionality, the *meaning* of, say, 'brown cow' is a function of "the meanings of 'brown' and 'cow' together with its syntax" (p. 334). But, by conceptual-role semantics, the *role* of 'brown cow' is a function of the roles of 'brown' and 'cow' *and* "*what you happen to believe about brown cows.* So, unlike meaning, inferential role is … *not* compositional" (p. 334). I take it that they conclude this because they take the role of 'brown cow' to depend on something *in addition to* the roles of 'brown' and 'cow'. But that doesn't seem to be the case: Granted, the role of 'brown cow' depends on the roles of 'brown' and 'cow'. What are those roles? Well, they *include* all of my beliefs that involve 'brown' and 'cow', and *that* includes my beliefs about brown cows. So nothing seems to be added. Now, there *is* a problem—the threat of circularity, viz., that, at bottom, the meaning of 'brown cow' will depend on the meaning of 'brown cow'—but that doesn't seem to be what Fodor and Lepore are complaining about at this point. Putting that aside for the moment, inferential role *does* seem to be compositional, so it *could* be what meaning is.

Earlier, however, we saw that the *meaning* of 'brown cow' has to be a constituent of a proposition—call such a constituent a "concept" for now. So we have two options: (1) identify propositions and concepts with roles, or (2) assert that there are two *kinds* of meaning: (a) a sentence means a proposition (and a sub-sentential expression corresponds to a concept), and (b) a sentence (or sub-sentential expression) means (or *is*) its role. Now, there's ample historical precedent for bipartite theories of meaning like (2). We might even think of propositional/conceptual meaning as a kind of referential meaning. Note that we would then have *three* kinds of referential meaning: classical Fregean *Bedeutung*, internal reference (as discussed above in §§2.7.1 and 2.8.2 and below in 8.3.1) and our new propositional/conceptual sort, which is not unlike a Meinongian theory of meaning (cf. Meinong 1904; Rapaport 1976, 1978, 1981, 1985/1986, 1991b, and references therein). Role meaning would be a kind of *Sinn*. One problem with such a theory is that it doesn't tell us what propositions or concepts *are*. That's an advantage to option (1), that of identifying propositions/concepts with roles (but cf. George Bealer's PRP theory **REF**, **[???]** which takes propositions as primitives). I won't take a stand on this now, though I lean towards the first option.

Fodor and Lepore's point is that if I believe that brown cows are dangerous but do not believe that being brown or being a cow is dangerous, then the concept of *dangerous* might be part of the role of 'brown cow', yet not be part of the roles of either 'brown' or 'cow'. Here, I think, it's possible that Fodor and Lepore's emphasis on *inferential* role rather than *conceptual* role might be misleading them. For me, being dangerous might be *inferrable* from being a brown cow without being inferrable from being brown or being a cow, *simpliciter* (that is, it's a sort of emergent property or merely contingently but universally true of brown cows). However, if being dangerous is part of the *conceptual role* of 'brown cow', it's also—*ipso facto*—part of the conceptual roles of 'brown' and 'cow'. It can't *help* but be. If *inferential* role, then, is *not* compositional, but *conceptual* role *is*, then so much the worse for *inferential* role. Inferential role, in any event, is subsumed by the broader notion of conceptual role. At most, then, Fodor and Lepore may have successfully shown why meaning (probably) isn't *inferential* role. Conceptual role, so far, emerges unscathed, despite Fodor and Lepore's claim that their argument is "robust ... [and] doesn't depend on ... how ... inferential role" is construed (p. 335). (Their argument does, however, appear to weaken Hartry Field's (1977) interpretation in terms of subjective probabilities.)

More, perhaps, needs to be said about compositionality. Let's look at it from the SNePS viewpoint. In SNePS, there are two kinds of nodes: *Molecular* nodes have structure, in the sense that they "dominate" other nodes; that is, a molecular node has one or more arcs emanating from it.[24] *Base* nodes, on the other hand, are structureless; that is, they do not dominaate any nodes, though they are dominated by other nodes. (An isolated base node would be a "bare particular" (Allaire 1963, 1965; **see also Nous 1: 211–212, 4: 109–134, 209–223**)
[???] or a "peg" (Landman 1986); but SNePS forbids them.) Following Woods 1975, we also distinguish between *structural* and *assertional* information about a node. Roughly, a node's structural information consists of the nodes it dominates; its assertional information consists of the propositional nodes that dominate it.

For example, consider the network of Figure 4.11. It contains 7 base nodes (`B1`, `B2`, `"John"`, `"rich"`, `"person"`, `"Mary"`, `"believe"`) and 11 molecular nodes (`M1`, ..., `M11`).[25] Consider `B1`: As a base node, it has no structure, hence no structural information, but we know several things about it assertionally: It (or, rather, that which it represents) is named 'John' (`M2`), it is rich (`M6`), and it is a person (`M4`). Consider `M4`: Structurally, it is (or represents) a proposition that `B1` is a person (that is, its constituents are `B1` and `M3`, the latter of which is (or represents) a concept

[24] A node *dominates* another node if there is a path of directed arcs from the first node to the second node.

[25] Hill 1994 would not consider sensory nodes (at the heads of `LEX` arcs) to be base nodes.

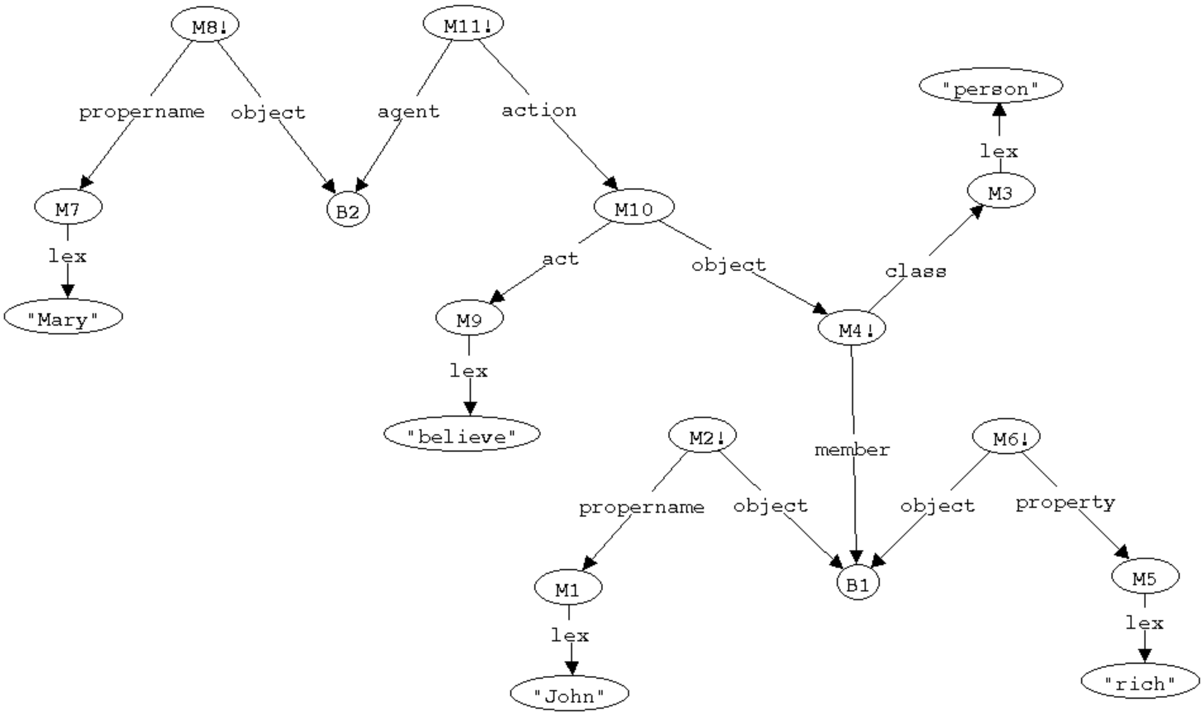


Figure 4.11: `M2!` = `B1` is named 'John'
`M6!` = `B1` is rich
`M4!` = `B1` is a person
`M8!` = `B2` is named 'Mary'
`M11!` = `B2` believes that `M4`

whose only structure is that it is lexicalized as 'person'). Assertionally, we know of `M4` that it is believed by Mary. (We also know, since it is an "asserted" node, that it is believed by Cassie; this, too, is probably part of its assertional information, even though it has nothing to do with node domination.)

Now, what does `M4` mean? *Structurally*, its meaning is determined by the meanings of `B1` and `M3`. For now, let's take the meaning of `B1` to be a primitive (or perhaps the node `B1` itself). The structural meaning of `M3` is determined by the meaning of the `"person"` node, which, again, we'll assume is either primitive or the node itself. So far, so good for compositionality. However, if meaning is conceptual role in the *entire* network, then we must also consider `M4`'s *assertional* meaning, which is that Mary (and possibly Cassie) believes it. Is assertional meaning compositional? This may be a matter of legislation. Let's suppose, however, that it is. Then the assertional meaning of `M4` is determined, let's say, by the assertional meaning of `M10` (which is the only node that directly dominates `M4`—ignore Cassie for now), which, in good compositional turn, is determined by the assertional meaning of `M14`. What's the assertional meaning of `M14`? As with base nodes, we could say that it is some sort of primitive or else the node itself. We could also say that at this point we must revert to *structural* meaning. That, in turn, suggests that for the *structural* meaning of a *base* node, we could revert to its *assertional* meaning. To make matters more complex, presumably the meaning of, for example, `M8` and `B2`, also play some role in the assertional meaning of `M4`.

I will leave for another time (and another researcher, Hill 1994) the spelling out of the details. But there are two observations to be made: (1) Circularity abounds. (2) Compositionality does *not* seem to be compromised. **[???]**[26] I might also note that productivity, systematicity, and isomorphism likewise do not seem to be compromised or rendered inexplicable. (We'll return to circularity.)

4.4.7.2.2 ***The analytic–synthetic distinction.*** What happened to the analytic–synthetic distinction? The proposal is to save inferential role by limiting it to *analytic* information: Analytic inferential role *is* compositional, so it *can* be identified with meaning. The first thing to notice is that this *removes* "being dangerous" from the meaning of 'brown cow' (and *a fortiori* from the meanings of 'brown' and 'cow'). Now, there are advantages and disadvantages to that. One of the disadvantages is that if I do believe that brown cows are dangerous, then that *is* part of the meaning of 'brown cow' (and my concept of brown cows is equally part of what 'dangerous' means to me). If, for example, the first time I read 'dangerous' is in the sentence 'brown cows are dangerous', then what 'dangerous' *means*, for me, is: something that brown cows are. Now, as we saw in §2.8.2, the more occurrences of 'dangerous' (or of 'brown cow') I encounter, the less likely it will be that 'brown', or 'cow', or 'brown cow' will play a significant role (excuse the expression) in my understanding of 'dangerous' (and, *mutatis mutandis*, the less likely it will be that 'dangerous' plays a signifcant role in my understanding of 'brown cow'). What will be left when such idiosyncratic, contingent aspects of the meaning play smaller and smaller roles (or drop out of my dictionary-like definition of 'brown cow' or of 'dangerous')? What will be left may well be just the analytic inferential roles: 'brown cow' will *mean* "cow that is brown" (although I might still *believe* that brown cows are dangerous, and have a connotation of danger whenever I encounter 'brown cow'). That's the *advantage* of analytic inferential role.

Of course, it's not *enough*. What about the meaning of 'cow' *tout court*? We have a few options even within the family of role-type semantics.

Option 1: 'cow' means "cow", where "cow" is a primitive term of Mentalese or of my language of thought (or a SNePS node). Perhaps this is what Fodor has in mind when he claims that we have innate concepts of, say, internal combustion engines **REF]**. Option 1 is OK as far as it goes, but not very enlightening.

Option 2: 'cow' means my entire set of concepts minus "cow", where "cow" is as in Option 1. That is, the meaning of 'cow' is its entire role (or location) in my entire mental network. That's holism. I think it's fine, as I argued earlier. But I grant that it seems to be a bit too much. So, when needed, we can choose Option 3:

Option 3: 'cow' means that portion of my entire set of concepts (minus "cow", of course) from which I can infer whatever else I need to know to use and understand 'cow'—that is, that more or less stable portion of my conceptual net that corresponds to the sort of information given in a dictionary or small encyclopedia. (This would be one implementation of the SCOPE mechanism of Hill 1994. Ehrlich 1995 limits SCOPE by, roughly, the information necessary to categorize the term.)

[26] Is this consistent with Hill 1994? Perhaps not.

What about circularity? Accepting—as I do—both compositionality and *conceptual*-role semantics (rather than mere *inferential*-role semantics, analytic or otherwise), we see that compositionality never "bottoms out". This, I take it, is one of the prices of the holism of conceptual-role semantics. How expensive is it? Well, note first that it rears its head when we inquire into the meanings of base nodes. Perhaps the structural–assertional distinction renders that head less ugly than it might otherwise appear. The other place that circularity appears is when we try to find a natural "stopping place" in the computation of a nodes' "full" (that is, both assertional and structural) meaning (cf. Quillian 1967, 1968). How bad is *that*? Don't forget: Our network is huge, and includes internal representations of all of the entities that a Harnad-like grounded theory postulates. We *could* say that the meaning of any node can never be given in isolation—to understand one node is to understand the entire network. We *could* say that the meaning of *some* nodes *is* intrinsic or primitive or given in some sense (Perlis (1991, 1994) seems to say this, and my treatment of Lakoff and Johnson (§3.2.2.2.3) is in a similar spirit). Or we *could* say that some smaller portion of the entire network is sufficient (this is the dictionary-like–definition strategy). We could *also* say all of the above, distinguishing different kinds of meaning for different purposes.

Fodor and Lepore aren't happy with analytic inferential role, however. First, the only way to identify the *analytic* inferences (from all the others) is to see which ones are validated by meanings alone, but the only way to identify meanings is to look at analytic inferences. I have no stake in defending analytic inferential role. I think that the notion of a broader *conceptual* role, limited at times as in Option 3, avoids this problem. Analytic inferences can be identified quite easily: They're the ones of the form Adj + N $\rightarrow$ Adj and Adj + N $\rightarrow$ N (more precisely albeit it still roughly, they're the ones of the form $\forall x[ANx \rightarrow Ax]$ and $\forall x[ANx \rightarrow Nx]$, where A is a predicate modifier). There are, of course, well-known problems with toy guns and small elephants, but even Fodor and Lepore are willing to waive these (p. 334).

Second, they see analytic inferential role as "jeopardizing" "the naturalizability of inferential role semantics" (p. 336), because it can't be identified with *causal* role, in turn because there is no causal theory of analyticity. I don't know what a causal theory of analyticity would look like. If it would be a theory explaining why we tend to infer N from AN (we do, after all, *tend* to think of toy guns as guns, and there is a sense in which small elephants *are* small, at least as far as elephants go), then I see no reason why we would even *want* to identify (analytic inferential) role with *causal* role. The former seems quite abstract and general; the latter seems to be a mere implementation of it, hence less interesting or theoretically important. And why naturalize semantics at all? Put otherwise, isn't it natural—and ubiquitous—to begin with?

4.4.7.3 The inconsistency.

So the inconsistency that Fodor and Lepore see in the compositionality/role/analytic–synthetic triad is this: If meaning is (inferential) role, then it is not compositional. If meaning is *analytic* inferential role, and if there is a viable analytic–synthetic distinction, then meaning *is* compositional. Moreover, analytic inferential-role semantics entails the analytic–synthetic distinction. But there is no viable analytic–synthetic distinction. There appear to be three options: (1) Keep compositionality and reject both the analytic–synthetic distinction and both inferential- and analytic-inferential–role semantics, (2) keep non-analytic inferential-role semantics and reject both the analytic–synthetic distinction and compositionality, and (3) reject all of them.[27] Of these, Fodor and Lepore ought to opt for (1).

Their first consideration is to resurrect the analytic–synthetic distinction in a limited form, namely, to allow it "only between expressions and their *syntactic constituents*" (p. 338). That's fine by me (see my discussion of AN $\rightarrow$ N and AN $\rightarrow$ A inferences). The problem with this that Fodor and Lepore see is that it rules out as analytic such statements as that cows are animals (or, presumably, that bachelors are unmarried men). That's fine by me, too, tradition be damned. Unless 'bachelor' is *defined* as 'unmarried man', it really *isn't* analytic that bachelors are unmarried men. A Martian sociologist trying to figure out what's "natural" about the category of bachelors would not treat the claim that bachelors are unmarried men as analytic (cf. Rapaport 1981, Lakoff 1987; see also the discussion of reverse engineering in Weizenbaum 1976, esp. p. 134). For Fodor and Lepore, that cows are animals must be analytic if what counts is *inferential* role. But, first, that has to be a rather broad definition of inference. And, second, it's just another reason for preferring *conceptual*-role semantics, which doesn't license any *analytic* or *logical* inferences from cow to animal. As Fodor and Lepore point out, "If Quine's arguments show anything, they show that there is no way to reconstruct the intuition that 'brown cow $\rightarrow$ animal' is definitional and 'brown cow $\rightarrow$ dangerous' isn't" (p. 339). I agree; but there *is* a way to distinguish these from the strictly definitional 'brown cow $\rightarrow$ brown', and that's all we need.

Their second consideration is that the holism of inferential-role semantics entails "that expressions in different languages are semantically incommensurable" (p. 339). Yes; so what? Does that prevent us from communicating—successfully—with one another? No—for reasons why, see Chapter 5. Ah—but *is* inferential-role semantics thus holistic? Fodor and Lepore think not: They think that the following argument is not a good one (p. 340):

1. The meaning of an expression is determined by *some* of its inferential relations.

2. "There is no principled distinction between those of its inferential relations that constitute the meaning of an expression, and those that don't" (p. 340).

3. $\therefore$ The meaning of an expression is determined by *all* of its inferential relations.

Premise 1 follows from inferential-role semantics, premise 2 follows from the *lack* of an analytic–synthetic distinction, and the conclusion is holism. They think that this is not a good way to argue for holism, because it is a slippery-slope argument *and* because it depends on denying the analytic–synthetic distinction. The latter is a problem because if you *accept* a principled analytic–synthetic distinction (as I do), you can't accept (2), and if you *deny* a principled analytic–synthetic distinction, you can't accept (1), because (1) requires a principled analytic–synthetic distinction. It seems to me that all that this shows is that holism can't be inferred this way, not that holism is false.

[27] Here's why: There are four principles: compositionality, the analytic–synthetic distinction, inferential-role semantics, and analytic-inferential–role semantics. So there are 16 possible combinations. Rejecting the analytic–synthetic distinction eliminates 8 of them (the ones in which the analytic–synthetic distinction is true). The analytic-inferential–role semantics $\rightarrow$ analytic–synthetic distinction relation eliminates another four (the ones in which analytic-inferential–role semantics is true but the analytic–synthetic distinction is false). Of the remaining 4, the inferential-role semantics $\rightarrow \neg$compositionality relation eliminates the one in which inferential-role semantics and compositionality are true.

Here's how I see it: (1) *is* true. In fact, I can give it at least two interpretations on *conceptual*-role semantics, not *inferential*-role semantics:

(1a) The *structural* meaning of an expression (or node) is determined by the expressions (or nodes) that constitute (or are dominated by) it.

(1b) The *dictionary-like* meaning of an expression (or node) is determined by *some* of its conceptual relations. (Which ones depend on the contexts in which the cognitive agent has encountered the expression and on which of those are needed to provide a "stable" meaning.)

Premise (2) *is* false. There are *lots* of *different* principled distinctions. One is that between *logical* inferences and *non*-logical ones (between ones whose logical form is AN $\rightarrow$ N or AN $\rightarrow$ A and ones whose logical form is $A \rightarrow B$). Another difference is that produced by (1a): the distinction between structural and assertional information. Yet another is that produced by (1b): the distinction between "core" relations and "peripheral" (or "connotational") ones. (I admit that the third has not been spelled out here. But Ehrlich 1995 sketches it out; the proof will be in the computational pudding.) Holism, as I see it, is independent of (1) and (2). But it *does* follow from—indeed, it simply *is*—the notion of the *full meaning* of an expression (or node) as given by conceptual-role semantics.

So the "crack in the foundations of" semantics (p. 342) can be patched by using different brands of role semantics, analytic–synthetic distinctions, and maybe compositionality: Buy *conceptual*-role semantics, a logical (or structural) analytic–synthetic distinction, and some version of compositionality—and accept that there are *lots* of aspects to "the" meaning of an expression.

4.5 HOW TO COMPARE ROLES.

One of the leftover problems that Fodor and Lepore saw has to do with the apparent incommensurability of different systems of roles. Perhaps, they suggest pessimistically, one will have to be reconciled to a theory of *similarity* of meaning, rather than of identity of meaning.

There are, I think, cases where roles indeed can't be cleanly compared. The clearest cases come from language translation. The role of the French preposition '*à*' is simply not played by any one preposition in English, nor is the role of the English preposition 'in' played by any one preposition in French. This prevents neither translation nor mutual comprehension. Cases of dissimilar roles among nouns also do not prevent everyday translation or comprehension, though they wreak havoc with literary and poetic translation, not to mention puns and even everyday associations or connotations. So be it. One can always convey the foreign meaning by a suitable, if prosaic and pedantic, gloss (cf. Rapaport 1981, Jennings 1985).

There are ways to compare roles "on the fly", though one has to look at the larger picture—indeed, larger and larger pictures—and one has to settle, sometimes, for only partial agreement. As Nicolas Goodman (personal communication) put it, you may recall, "... I associate with your words various complexes of memory, behavior, affect, etc., in such a way that I end up with a sentence which can play *more or less* the same role *in my life* as your sentence plays *in your life*" (my italics). The important point is that this correspondence (hence, this *semantic* understanding) *can* be set up. As Lenat and Feigenbaum (1991) observe about a similar situation, "While this does not guarantee that the genuine meanings of the concepts have been captured, it's good enough for us" (p. 236). What is "genuine meaning"? Is it an "intended interpretation"? Intended by whom? In the case of Lenat and Feigenbaum's CYC system—a vast, encyclopedic knowledge base (but one that can be thought of as akin to the mind of a (computational) cognitive agent; cf., however, Smith 1991)—there is an answer: The genuine meaning of a concept is the one intended by the CYC researchers. But in the case of a human or of a CYC-like system that "changes its mind" and "learns", *its own* understanding is just syntactic. More importantly for our present concern,

> After all, how does one guarantee that one's neighbor shares the same meanings for terms? The answer is that one doesn't, at least not formally or exhaustively. Rather, in practice, one defeasibly assumes by default that everyone agrees, but one keeps in reserve the ubiquitous conflict resolution method that says "one may call into question whether they and their neighbor are simply disagreeing over the meaning of some terms". (Lenat & Feigenbaum 1991: 236.)

This is the issue we take up next: How does communicative negotiation enable us to understand one another?

Chapter 5

COMMUNICATION, NEGOTIATION, AND INTERPRETATION

> They'd entered the common life of words. ...
>
> After all, hadn't the author of this book turned his thoughts into words, in the act of writing it, knowing his readers would decode them as they read, making thoughts of them again? (Barker 1987: 367.)

> A book is a way to hold the mind of another in your hands. ... [???] Books. How you reach across time and space to be with another's mind. (Advertisement for Doubleday Book Stores, *The New Yorker* 1991 [???], p. 112.)

> Researchers concerned with modeling people recognize that people cannot be assumed to ever attribute precisely identical semantics to a language. However, the counterargument is that computers *can* be programmed to have precisely identical semantics (so long as they cannot modify themselves). Moreover, as evidenced in human coordination, identical semantics is not critical, so long as satisfactory coordination can arise. (Durfee 1992: 859.)

> When you and I speak or write to each other, the most we can hope for is a sort of incremental approach toward agreement, toaward communication, toward common usage of terms. (Lenat 1995: 45.)

5.1 COMMUNICATION.

I have placed a fairly heavy burden on the role of communication. For it is there that I have swept all the problems left over from our discussions of misunderstanding and conceptual-role semantics

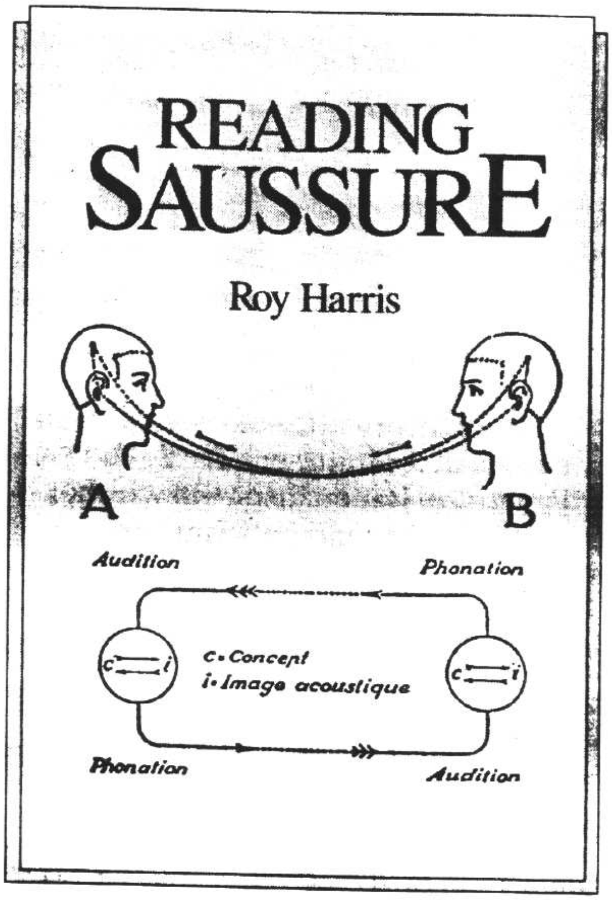


Figure 5.1: My ideas travel from my mind, to my mouth, to your ears, to your mind, and conversely. (From the dust jacket of Harris 1987.)

(cf. §§2.7.1, 3.2.2.2.4, 3.2.2.2.5, 4.4.2.1, 4.4.6, 4.4.7.3). It is through the process of interactively (i.e., reciprocally) communicating with another that cognitive agents come to learn language, to correct misunderstandings, and to change one another's minds. Such communication allows one to "align" one's own knowledge base, expressed in one's own language of thought, with another's. In this chapter, we will explore this idea in more detail.

As expressed in the opening quotations, a standard way of looking at communication is that the only way for me to know what's going on in your mind is for you to express your ideas in language—to "implement" them in words, say—and for me to translate from that "public communication language" (Shapiro 1993) into my own ideas. An example is the miraculous, magical mystery of reading: When we read, we seemingly stare at a bunch of arcane marks on paper, and suddenly come to know of events elsewhere in (or out!) of space and time. How? By having an algorithm that maps the marks (which have a syntax) onto our concepts, i.e., by interpreting the marks. Conversely, in speaking, my ideas travel from my mind, to my mouth, to your ears, to your mind, as in Figures 5.1 and 5.2.

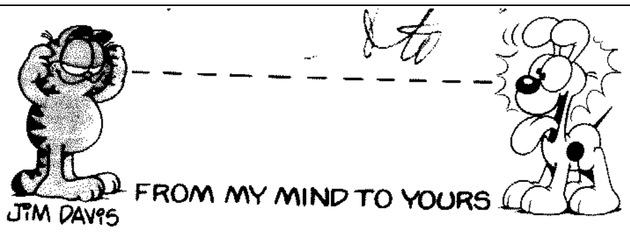


Figure 5.2: Information gets sent from the sender's (or Garfield's) mind to the recipient's (or Odie's) by being written on paper (that is, by being implemented in language). (From Post-It Note P-788, ©1978, United Feature Syndicate.)

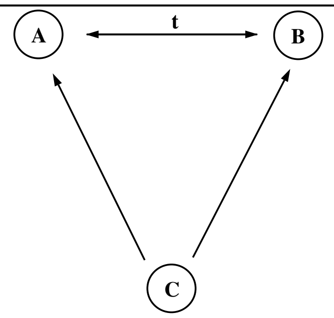


Figure 5.3: A and B are cognitive agents communicating about C, a real object in the external world. The arrow between A and B represents the communication between them of some term t expressing C. The arrows from C to A and to B represent A's and B's (joint) sensory access to C.

What is it, however, that we are talking about? It would seem, from the simplified picture of Figure 5.1, that we are only talking about our own ideas. What about the real world? Surely, we often talk about some external object that the two of us have joint access to. Isn't that, after all, how we know that we're talking about the same thing? Isn't the picture really as in Figure 5.3? In Figure 5.3, the idea is that two cognitive agents A and B use some term t to refer to some external object C in the real world. Both A and B have independent, direct access to C, and so can adjust their understanding of t by comparing what the other says about C with C itself. Is that not how things work?

As usual, the answer is: Yes and No. The picture is still too simple, even though it is already an elaboration of Figure 5.1. What's missing is that A's access to C results in a private, internal idea (or set of ideas) about C, and similarly for B. Further, *it is these private, internal ideas* that A and B are talking about, *not* C. So the picture is more like the Rube-Goldbergian Figure 5.4. Here, cognitive agent A perceives external object C and constructs (or finds) her own mental representation of C; call it C_A. A then wishes to inform cognitive agent B of what she (A) is thinking, and so utters t, some linguistic expression that Fregeanly denotes C and internally means C_A. Cognitive agent B hears t and constructs (or finds) his own mental representation. Here, there are two possibilities: (1) B takes A to be talking about what B thinks of as C_B, namely, B's own mental representation of C; this is close to the ideal situation, the case of perfect mutual

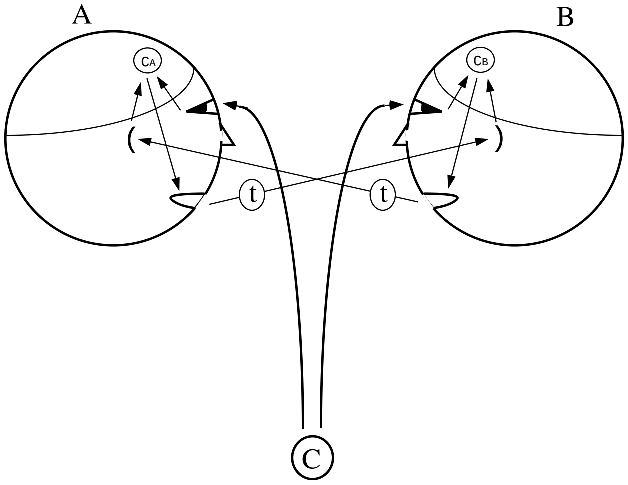


Figure 5.4: A perceives C, constructs mental representation C_A of C and utters t; B hears t and constructs mental representation C_B. Similarly, B perceives C, constructs C_B, and utters t; A hears t and constructs C_A. (See text.)

understanding. (2) B takes A to be talking about something distinct from C_B; this is the case of miscommunication, of misunderstanding. In case (1), B comes to believe that A is thinking of the "same" thing that he (B) is thinking of. B could continue the conversation by saying something else about C_B, using t. A hears t and constructs (or finds) her mental representation in one of two ways, just as B did. Again, then, we have either a case of perfect understanding or a case of misunderstanding. In case (2), where B has misunderstood A, B might say something that will alert A to the misunderstanding. By continued communication, A and B will negotiate about what it is they are talking about, hopefully coming to some agreement.

In Figure 5.4, the arrows between A and B represent the attempt at communication of some term expressing an object. The arrows from A's eyes and ears to C_A and from B's to C_B represent their individualized, private access to the *perspectival* objects C_A and C_B. Those objects can be thought of as (representing) Meinongian objects or Castañedian guises or propositions. The arrows from C to A and to B represent the causal connections between C and A and between C and B. C itself is—from the first-person points of view of A and of B—a hypothetical or assumed real object, accessed only indirectly via C_A and C_B. The linguistic expression t, to use Oliver Sacks's phrase, is "symbolic currency" used "to exchange meaning" (Sacks 1990a, quoted in Sacks 1990b: 3). Expressions like t constitute the text of the "books" that enable us "to be with another's mind". (But even this is not quite right, as we will see in §5.4.)

5.2 NEGOTIATION.

Interpretations are negotiated in interaction. Every time we talk, we negotiate interpretations about referential and social meanings. The more intense and frequent the interaction between speakers with diverging interpretations of the meanings of a word, the more likely a 'negotiated settlement' will obtain, more or less spontaneously, through

> linguistic usage. When interpretations become conventionalized, we call that 'meaning'. But even so, that new meaning is subject to revision and negotiation. (Alvarez 1990.)

So, negotiation is the key to understanding. Candace Sidner (1994) points out that discourses among collaborators function as negotiations and that discourses containing negotiations serve to establish mutual beliefs. Miscommunication (case (2), above) is in fact the norm. Suppose that I think of C_B when I hear you utter t (as a result of your thinking of C_A, which in turn is caused by your perception of C). Even if this C_B is the C_B that I think of when I *perceive* C, still, C_B will play a different role in my network than C_A does in yours. So how do we understand each other, as—apparently, or for all practical purposes—we do?

First, why is there a problem at all? Why is there the potential for (and usually the actuality of) miscommunication resulting in misunderstanding? The answer is simple:

> . . . transmission of representations themselves is impossible. I cannot be sure that the meaning of a word I say is the same for the person to whom I direct it. Consequently, language works as a system of values, of reciprocal expectations. To say it differently, the processes of verbal communication always constitute a try, a hypothesis, and an intention from the sender to the receiver. (Vauclair 1990: 321–322.)

On this view, if you could literally read my mind (as, indeed, I *can* literally read Cassie's; cf. Carnap 1956: 244–247; Simon 1992: 6–7), there would be no misunderstanding, hence no miscommunication. But, since you can't, there is (cf. Fig. 5.5). Arguably, though, even this wouldn't suffice. For you would still have to understand my language of thought, just as a reader of a text written in one's native language must interpret that text even though the language is common. So it's highly unlikely, except possibly in the most artificial of situations (as with Cassie) that communication can ever be "perfect".

Fortunately, as we noted in §§2.8.2 and 4.4.6, the process of language understanding is self-correcting. This does not *guarantee* mutual understanding, but (a) it makes it very much more likely, and (b) it makes residual misunderstandings of marginal relevance. The latter happens in two ways. First, suppose for the sake of argument that Cassie's and Oscar's mental networks differ by only three nodes. Suppose, for example, that Cassie believes some simple, 3-node,`OBJECT-PROPERTY`

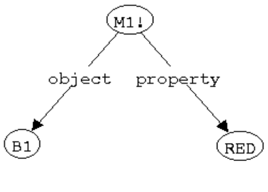
Figure 5.5: How misunderstanding can arise.

proposition[1] about an object but that Oscar doesn't. (This is, admittedly, implausible for all but the case of toy computational cognitive agents such as Cassie and Oscar, but it will serve to make my point.) Then, if Cassie tells Oscar something about some object, Oscar will not fully appreciate all the connotations of Cassie's claim, because her claim will be linked to the 3-node proposition that Oscar lacks. But what Oscar misses will, in general, be irrelevant. It will be of only marginal concern.

Second, suppose, again for the sake of argument, that Cassie's and Oscar's mental networks differ only in some of their `lex` nodes; that is, they are structurally the same. Suppose, for example, that Cassie is discussing chess, but Oscar is discussing a battle in the Civil War, or that Cassie is discussing mathematical lattice theory but that Oscar is discussing chemistry. As long as the two agents' interpretations of each other's utterances are isomorphic, neither will be able to determine that they are not talking about the same thing. Oscar, for instance, might not have the "intended interpretation" of Cassie's utterances; but this will make no practical difference:

> Jan and Edwige never understood each other, yet they always agreed. Each interpreted the other's words in his own way, and they lived in perfect harmony, the perfect solidarity of perfect mutual misunderstanding. (Kundera 1978: 227.)

[1]There are 3 nodes: one (e.g., `B1`) for the object, one (e.g., `RED`) for the property, and one (e.g., `M1`) for the proposition:



Following Lynne Rudder Baker (personal communication, 21 April 1989), let's call these 'crazy interpretations'. *Perhaps* Cassie and Oscar could calibrate their interpretations by reference to the real world. But I argue that this is not accessible to them (see the next paragraph but one). Any apparent such access is all internal. Hence, I cannot rule out crazy interpretations. But what makes such crazy interpretations irrelevant is the need for successful communication. Cassie and Oscar exist in a social environment, which constrains (or helps to constrain) the possible interpretations, even though it cannot rule out such "inverted spectrum" cases as where Cassie might be talking about a mathematical lattice and Oscar might understand her to be talking about the chemical structure of some molecule. Because they share a social environment, these differences will be irrelevant insofar as they have no pragmatic implications.

How, then, does negotiation work? How are mistakes detected and corrected? By a continual process of learning, hypothesis testing, and belief revision. The more we communicate with each other, the more we learn. We can ask questions and match the actual answer with our hypothesized one, or we can make trial statements and match our interlocutor's response with one we expect. If the question is not answered as we expect, or if the reply is surprising, we revise our beliefs. (Cf. Hirst et al. 1993.) By successive approximation, we can asymptotically approach mutual comprehension (cf. Rapaport 1976: 178–180, Rapaport 1985/1986: 84–85).

Communication between cognitive agents is not the only way to correct misunderstandings. Perception is another (cf. Maida & Shapiro 1982: 300–301). Both, however, are kinds of "communication" with something external to the understander. Crucially, both work in the same way: The understander compares two *internal* representations, one causally produced from the speaker or the act of perception, the other part of the antecedently existing internal mental network. When there is a mismatch, the understander must change his or her (or its) mind. "As long as the conversation proceeds without our getting into ... [a] situation [in which "we *didn't* know what was meant"], the system [i.e., the cognitive agent] has all the connections with reality it needs" (Shapiro & Rapaport 1987: 271).

What enables this self-correcting ability is the same thing that enables a computer to understand natural language, namely, the fact that—as we saw in §2.7.1—"computers, like us, participate in the real world: they take real actions" (Smith 1985: 638). Those actions affect us, and, conversely, ours affect them. We are, thus, in the same social environment, subject to mutual correction. We're all in the same boat.

5.3 BRUNER'S THEORY.

Jerome Bruner's studies of language acquisition shed light on communication and negotiation. According to Bruner, children *interpret* and *negotiate* during their acquisition of language:

> The negotiation [between adult and child] has to do, probably, least with syntax, somewhat more with the semantic scope of the child's lexicon, and a very great deal with helping make intentions clear and making their expression fit the conditions and requirements of the "speech community," i.e., the culture. ... The development of language ... involves two people negotiating. ... If there is a Language Acquisition Device [LAD], the input to it is not a shower of spoken language but a highly interactive affair shaped ... by some sort of an adult Language Acquisition Support System [LASS].

(Bruner 1983: 38–39.)

In a passage that is virtually a summary of much that I have been urging, Bruner sets out an example of language acquisition by the child:

> ... reference can *vary* in precision from a rather wooly vagueness to a proper singular, definite referring expression. Indeed, two parties to a conversation may refer to the "same" topic with widely different degrees of precision. The "electricity" that a physicist mother has in mind will not be the "same" as what her child comprehends when she warns him about getting a shock. Still the two may carry on about "electricity" in spite of this indefiniteness. Their conversational negotiation may even *increase* her child's definiteness. Truth is not all that is involved in such causal chains. The child's conception of electricity may be vacuous or even wrong, yet there is a joint referent that not only exists in such asymmetric conversations, but that can be developed both for its truth value and its definiteness. (Bruner 1983: 67–68.)

Some observations are in order. By 'reference', Bruner must mean the *act* of referring, for reference as understood, say, in the Fregean way is an all-or-nothing affair: A word either refers or it doesn't. But *acts* of referring *could* "vary in precision"—speakers can be more or less careful, more or less sloppy, more or less detailed in their use of words.

The fact that Bruner chose to use scare quotes when stating that the two speakers "may refer to the 'same' topic" suggests that, indeed, "the" topic is *not* the "same", or else that the referring expressions used are associated with "widely different" concepts. So, again, he is not talking about the Fregean referent of the word. The physicist mother and her child have "widely different" concepts associated with 'electricity'. Indeed, the physicist will have a vast, complex network of meaning, vaster still than the ordinary adult, whereas initially the child will have none (it will be "vacuous") (or, at best, the child will have a concept of something—he or she knows not what—called 'electricity'). What *is* the "same" is the *referring term*; the associated (mental) concepts are *different*. There is no place (so far) in Bruner's description for the referent—electricity—itself. So, when mother and child "carry on about 'electricity' ", are they both talking about electricity itself? No, or not necessarily. "Truth is not all that is involved in such causal chains." Rather, they are talking "about" the word. Here, one must be careful not to confuse use with mention: *The only thing in common is the word.* There are *two*, *distinct* things that the word means: the physicist's meaning and the child's meaning. The goal—in the long term—is for the child's meaning to be as much like the physicist's as makes no difference. (This may in fact be too much to ask. As I noted, most parents are not physicists. So the goal need only be for the child's meaning to be as much like an ordinary adult's meaning as makes no difference.) As the "conversational negotiation" continues, the child's concept will become more detailed, approaching that of the mother.

"Truth is not *all* that is involved", but is it involved at all? Bruner does say, at the end, that "there is a joint referent", but what is that joint referent? It is surely not electricity itself, because the joint referent "can be developed ... for its truth value and its definiteness": Electricity itself has no truth value; 'electricity' does (well, it has a Fregean *referent*, to be more precise). Nor is electricity "definite" or "indefinite"; only our theories or *concepts of* electricity can be.

Could the joint referent be the *common concept* of electricity that the mother hopes will be established? If so, why should Bruner—or we—think there must be such a thing? What, indeed, would it be? What is in common is only the *word*; there are two *different* mental concepts associated with it, and the *child's* "can be developed". As we saw with Potts (§4.4.4), there is no need for a common external object. Rather, the picture we get from Bruner's description is this: Each person uses the *same word* to "refer" to his or her *own concept*; by negotiation, the concepts come into alignment.

There *is*, in a sense, *something* shared. Bruner says that "the means [for referring, or perhaps for the intent to refer] comprise the set of procedures by which two people establish 'jointness' in their attention" (Bruner 1983: 68). In what sense is their attention "joint"? Perhaps in the sense that what *I* am thinking of is what *you* are thinking of, though the 'is' here need not be the "is" of identity—it is more likely the "is" of equivalence or correspondence, as in Figure 5.6. What is "shared" is all in the *child's* mind (or all in the mother's)—shared by virtue of the child (or the mother) having *both* his or her *own* concept *as well as* his or her own *representation of* the *mother's* concept, plus some way of comparing them.

So, there is no need either for a joint *external* referent or for a joint *internal* referent. There is only need for sufficient similarity of structure of each conversant's internal networks for conversation to continue successfully:

> Achieving the *goal* of referring has little to do with agreement about a singular definite referent. It is enough that the parties to a referential exchange know that they share enough overlap in their focal attention to make it worthwhile continuing .... When the physicist mother tells her four-year-old that he has just been shocked by "electricity," she does not and need not assume that he has either the same extension or intension of the concept as she does. Nor need she care, if the conversation can only continue.
>
> The problem of how reference develops can, accordingly, be restated as the problem of how people manage and direct each other's attention by linguistic means. (Bruner 1983: 68.)

We will return to this notion of a speaker directing the interlocutor's attention in §9.4, when we look into non-human language use. For now, note that the picture we have, both from our own theory and from Bruner's, is this:

- A cognitive agent A refers to (communicates a reference to?) some object O (or some mental concept O_M) via term t to another cognitive agent B iff A directs B's attention to think about (for example, to find or else build in B's mental semantic network) an O or (O_M) concept expressed by t (for example, with a `lex` arc to t).
- Moreover, A's O_M and A's representation of B's O_M will be more or less equivalent, and B's O_M and B's representation of A's O_M will be more or less equivalent, where the equivalence becomes "more" rather than "less" by negotiation.

I'll conclude this discussion of Bruner (there will be more later, in §9.6) with one more quotation:

> John Lyons ... entitles an essay "Deixis as the Source of Reference." ... I think an equally strong case [can] ... be made ... that discourse and dialogue are also the

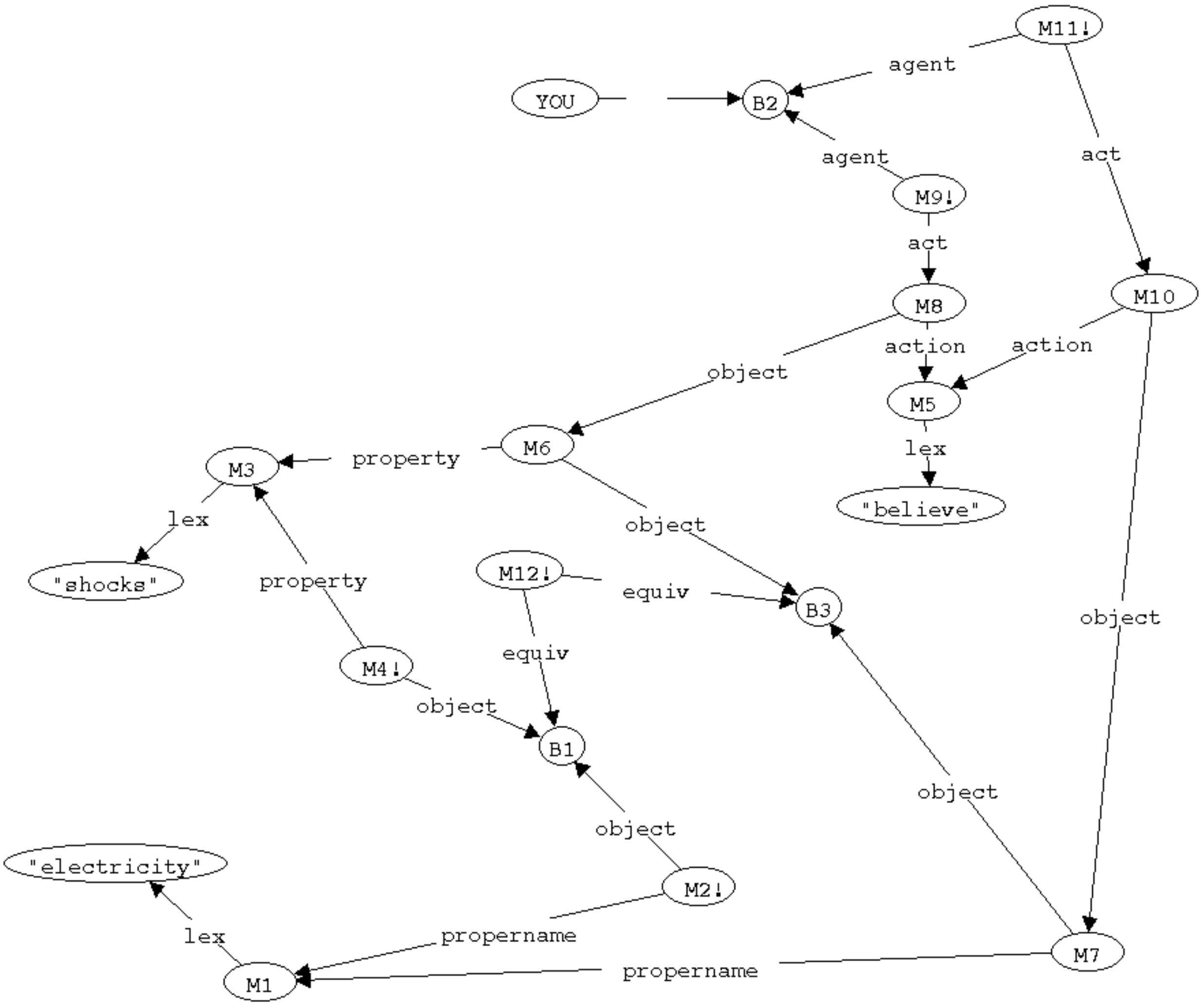


Figure 5.6: `M2!` = (I believe that) `B1` is called 'electricity';

`M4!` = (I believe that) `B1` shocks (i.e., I believe that there is something (viz., `B1`) called 'electricity' and that it shocks).

`M7` = `B3` is called 'electricity';

`M6` = `B3` shocks;

`M9!` = (I believe that) `B2` believes `M6` and `M7` (i.e., I believe that *you* (viz., `B2`) believe that there is something (viz., `B3`) called 'electricity' and that it shocks).

`M10!` = (I believe that) `B3` = `B1` (i.e., I believe that what you call 'electricity' is what I call 'electricity').

(The "you"-pointer mechanism is based on the I-pointer of Rapaport, Shapiro, & Wiebe forthcoming.)

> sources of reference. If they were not, each speaker would be locked in a web of isolated referential triangles of his [sic] own making—if indeed he could construct such a web on his own. (Bruner 1983: 88.)

That is, *negotiation* is a source of reference. More precisely, one way to get out of the "web of isolated referential triangles" (circles?)—to "ground" one's symbols—is by means of dialogue. Note, however, that even on Bruner's own view, this does not really get us "out" of our internal network, since all it can do is set up correspondences between objects in two belief spaces (together with correspondences with internal representations of external objects).

5.4 UNDERSTANDING AND GENERATING.

What happens in communication? When I speak—when I *generate* an utterance—I generate expressions "that are pertinent to the [neurophysiological] stimulus and are usable to narrate the primary [neurophysiological] display when inserted in appropriate grammatical structures" (Damasio 1989b: 25). For example, I perceive an object, which causes neuronal activity representing features and structure. These are linked to other neuronal structures that "generate names" (Damasio 1989b: 25) that, in turn, allow me to communicate two things to you: (1) *that* I am thinking of an object and (2) *what* it is.

But "I cannot be sure that the meaning of a word I say is the same for the person to whom I direct it" (Vauclair 1990: 321). Thus, symbols don't "*convey* meaning" (in the sense of the Calvin and Hobbes cartoon (Fig. 3.1). Rather, they *elicit* it in the mind of the receiver. With luck and negotiation, the ideas elicited in the receiver's mind correspond (because of structural similarity) to the same things that the speaker's symbols correspond to. The speaker's symbols act as stimuli to "activate" concepts in the receiver's mind. As my colleague Jorge J. E. Gracia has expressed it,

> We do not perceive ideas; what we perceive are certain phenomena that suggest to us certain ideas. If I ask you, for example, "Do you approve of what the President did?" and you frown in return, I conclude that you do not. But it is altogether possible that you do in fact approve ..., although you ... mislead me by making the frown. My conclusion that you do not, then, can be taken only as an interpretation of what you are thinking based on certain empirical evidence that is only indirectly related to what you think. (Gracia 1990: 495.)

I can't have direct access to your thoughts, only to your speech acts (to your language, including gestures). My interpretation is not of what you are thinking, but of your language and gestures. It is, indeed, a conclusion; understanding involves inference, albeit *defeasible* inference.

5.5 WINSTON'S PROBLEM.

Winston's Problem, recall (from §3.2.2.2.4), concerns what might happen if the knowledge-representation language of a computer system that can learn concepts (that is, its language of thought) differs significantly from that of humans. According to Winston, what would happen is that the two systems—computer and human—would not be able to understand each other. How serious is this problem? According to Joseph Weizenbaum, the intelligence of computers "must always be an intelligence *alien* to genuine human problems and concerns" (1976: 213). There are, it seems to me, several levels of difficulty:

1. Consider two cognitive agents, Cassie and Oscar, who share both a public communication language (say, English) and a language of thought. For concreteness, suppose their language of thought is the SNePS/Cassie knowledge-representation language (as described in Shapiro & Rapaport 1987). Winston's Problem would arise here only to the extent that it arises for any of us in everyday life: Insofar as our *experiences* differ—insofar as we have different background or "world" knowledge—then to that extent will we mutually misunderstand each other. As we have seen, though, the more we communicate, the more we will come to understand each other.

2. If Cassie and Oscar share only a language of thought, but *not* a public communication language, then there is an extra layer of difficulty due to the difficulties of translation. Still, with enough work and dialogue, this can be overcome.

3. In either of the above cases, things would be made worse if Cassie's and Oscar's "conceptual schemes" differ. By this, I don't mean that their languages of thought differ, but that their "world knowledge" is so different that even common experiences would be differently interpreted—and radically so:

   > The falling of a maple leaf is a sign of autumn ... because *we* have established a connection between them on the basis of certain observations and, therefore, use the phenomena in question to indicate something of interest to us. A different culture ... might see the falling of a maple leaf ... as [a] sign of other events or even as [an] indication of the divine will to punish and reward them. (Gracia 1990: 502; my italics.)

   This is not unlike the situation where there is a single computer program with two distinct input–output encodings, so that one computer is taken to be discussing chess while the other is taken to be discussing a Civil War battle (or one is taken to be discussing chemistry, the other, mathematical lattice theory). And anthropologists tell us that where Western physicians see viruses and bacteria, other cultures, such as the Kulina of Brazil, see *dori*—a substance "that permeates the flesh of shamans, giving them the ability to cure as well as to injure others"—injected into the body of a victim by a shaman (Pollock 1994: 18). Here things begin to get a bit more difficult. Nonetheless, it appears that we can *understand* the other's point of view, even if we disagree with it.

4. Winston's Problem becomes more threatening, of course, when the languages of thought differ. Even here, there are degrees of difference. For instance, Cassie and Oscar might both have SNePS languages of thought, but Cassie's might use the case frames that Shapiro and I advocate (Shapiro & Rapaport 1987) whereas Oscar might use those advocated by Richard Wyatt (1989, 1990, 1993). Here we have an empirically testable hypothesis that, say, one of the languages of thought would be "better" than the other in the sense of enabling the cognitive agent whose language of thought it is to understand finer discriminations. Conceivably, one of the languages of thought might be so (relatively) impoverished that

its "user" would simply not be able to understand or express some distinction that the other could.

5. Another level of difficulty—equally empirically testable—would arise if the two languages of thought were distinct members of the same general *kind* of knowledge-representation language. For instance, both might be symbolic, intensional, knowledge-representation and reasoning systems, say, (some version of) SNePS and (some version of) KL-ONE (cf. §1.2.5).

6. The *potential* for more serious inability to communicate occurs when one of the cognitive agents has a *connectionist* language of thought while the other has a "classical" symbolic one. This, I take it, is the situation Winston had in mind, though he wrote before connectionism was as popular as it is now. There would indeed be a problem if Cassie, whose language of thought was symbolic, tried to "read the mind" of Oscar, whose language of thought was connectionist. But as long as they spoke a common, public communication language, negotiation via dialogue might overcome any residual problems (cf. Hirst et al. 1993). This, too, is testable. (Indeed, for all we know, we are testing just such hypotheses as this and the ones in (4) and (5) every day when we speak!)

7. The worst case would be the Black Cloud case (§3.2.2.2.4): Here there is no common *kind* of language of thought, no common conceptual scheme, no common public communication language. This would appear to be a case for despair, though some are optimistic (e.g., Sagan 1980: 287–289). The optimism, note, comes from the hope that there is *enough* of a common basis to get negotiational dialogue off to a start.

5.6 SUMMARY.

When we communicate, we attempt to convey our internal meanings to an audience (interlocutor or reader) by means of a public communication language: "A book is a way to hold the mind of another in your hands." In so doing, however, we almost always fail. But we almost always nearly succeed. The resulting misunderstandings if near enough, can be ignored. But if we can't ignore them, we can minimize them through negotiation—learning what our audience meant or thought that we meant.

But minds are abstract. To be able to be in causal communication with them, they need to be implemented. Or, rather, ideas need to be implemented—to be expressed in a syntactic medium that can subsequently be (re-)interpreted in, or by, another mind. We will examine *implementation* in Chapter 7. First, however, there are some loose ends to tie up from our discussion of conceptual-role semantics and the first-person point of view.

Chapter 6

METHODOLOGICAL SOLIPSISM, INTERNALISM, AND THE FIRST-PERSON POINT OF VIEW.

6.1 INTRODUCTION.

Methodological solipsism (Putnam 1975, Fodor 19xx) [???] is the view that to understand the "psychology" of a cognitive agent, it is not necessary to specify the details of the external world in which the agent is situated and which impinge on the agent's sense organs. This is not to deny that there *is* such a world or that there is such sensory input—hence the qualifier 'methodological'. Rather, it is to acknowledge (or assume) that all that is of interest psychologically or cognitively can be studied from the surface inwards, so to speak (cf. Fig. 6.1. That is, to use Hilary Putnam's Twin-Earth example, it makes no difference psychologically or cognitively whether what I see or taste is chemically H_2O or XYZ. It only matters that what I see and taste is (say) cool, clear, tasteless, used for drinking and washing, etc., and how my mental representations of that fit in my mental semantic network.

We have already seen one argument for methodological solipsism in language understanding in our discussion of Potts (§4.4.4). And, as we noted in §5.3, Bruner's theory of language acquisition certainly seems compatible with methodological solipsism. In this chapter, we'll look at some other arguments and considerations in its favor, as well as some objections to it, and we'll emphasize its importance when dealing with the first-person point of view.

6.2 INTERNALISM.

Even if there is an external world (or, more conservatively, *despite* the existence of the external world), we deal with it *internally*. It plays no role in explaining the mind, because all it provides are *external* objects that are immediately *internalized* by our sense organs. On Twin Earth, XYZ is internalized (or mapped) to the "same" mental concept, viz., "water", that H_2O is on Earth—at least, "same" modulo conceptual role. I *talk* of water, I *think* of water, etc., and the same holds for Twin-me. We *drink* different stuff; but that's a *bodily* phenomenon—an *implementation* or *input*–

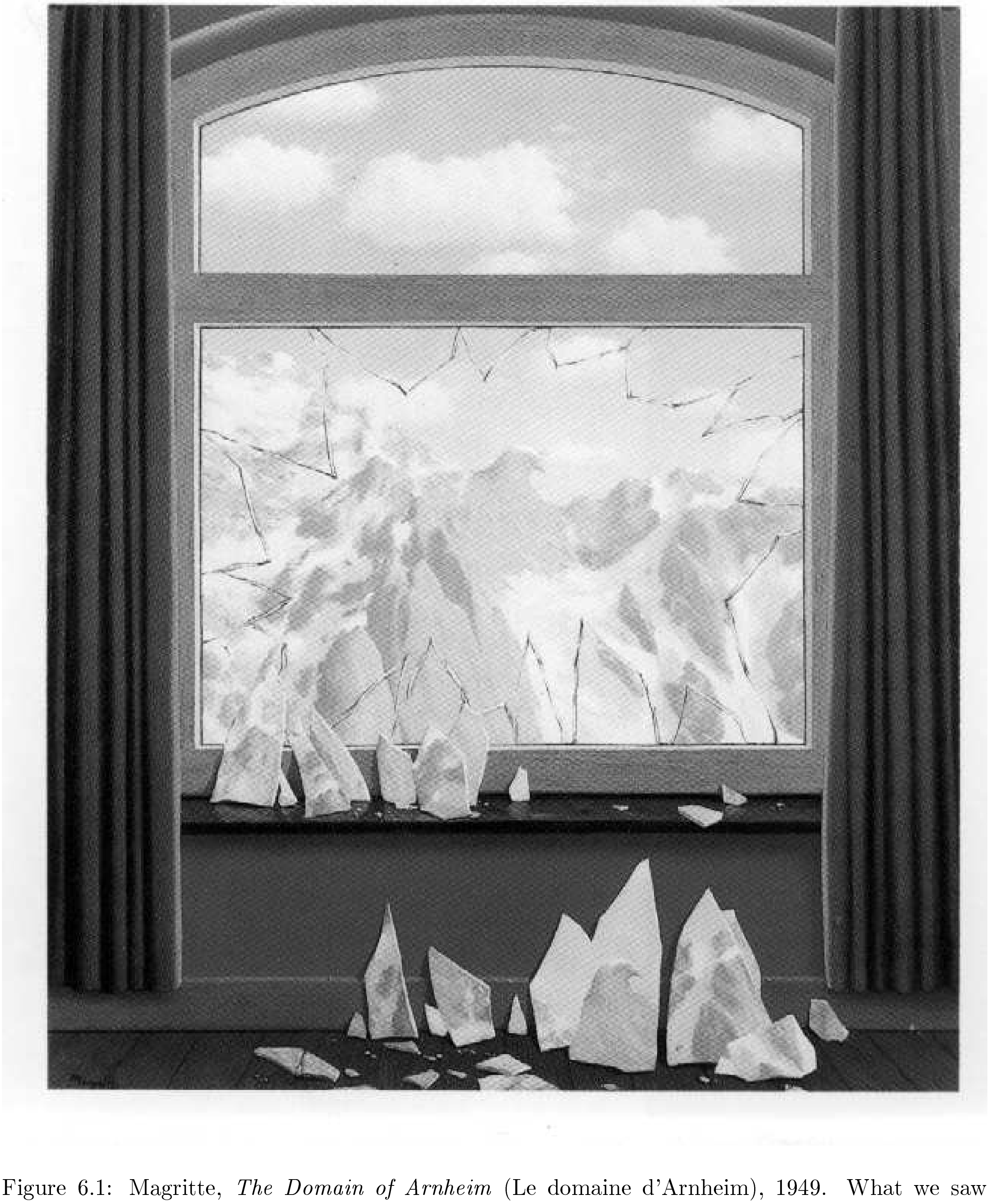


Figure 6.1: Magritte, *The Domain of Arnheim* (Le domaine d'Arnheim), 1949. What we saw "through" the now-shattered glass *was* painted on it, but *did* match the external world.

output phenomenon—not a *mental* phenomenon. We might say that drinking (washing, etc.) is a *de re* phenomenon, while talking, thinking, etc., are *de dicto* phenomena. And the mind, as I have suggested elsewhere, is essentially *de dicto* (Rapaport 1986a; Rapaport, Shapiro, & Wiebe 1986; Wiebe & Rapaport 1986; Shapiro & Rapaport 1991).

Aren't my beliefs *about* H_2O, though? They're not (or are they?) *about* XYZ or *about* my internal concept "water". This is difficult to answer, in part because the best answer is, no doubt, a legislation about how to understand 'about'. Let's say this, then: When I talk or think about water (e.g., when I say "Water is wet"), I *attempt* to talk about the stuff in the external world—H_2O—and I do so by uttering terms that express my internal concept "water". (I am, of course, not talking about my internal concept in the sense I would be if I said "My internal 'water' concept has 10 nodes dominating it.") Consider Helen Keller who, at the well house, "knew then that 'w-a-t-e-r' meant the wonderful cool something that was flowing over my hand" (Keller 1905: 36). Note that she did *not* say that 'w-a-t-e-r' meant H_2O, or XYZ. Twin-Helen would have had the *same* experience, and 'w-a-t-e-r' would have meant exactly the same thing for her (even though, as a matter of mere external *fact*, it would have been XYZ that was flowing over her hand).

To complicate matters, there are at least three things that my utterances and thoughts could be "about": my internal nodes, an abstract concept, or the external stuff out there. Ultimately, I am trying to "get at" the external stuff out there. I do so by using terms of a public communication language that express the internal nodes of my language of thought. Those internal nodes are my mental implementation of an abstract concept. There is a parallel case for propositional nodes, propositions, and states of affairs. My belief, say, that the pen with which I drafted this chapter is black is a belief "about" (in one sense) a certain state of affairs in the external world. I say it in English (as 'the pen with which I drafted this chapter is black'), and this English sentence expresses an internal propositional node of my language of thought, which is my mental implementation of a proposition. If I believed that a unicorn was in my garden yesterday, then there would (alas) be no corresponding state of affairs, though there *would* be a propositional node and a proposition. This is an old story by now, told in many different ways. (My version is in Rapaport 1976, 1978, 1981, 1985/1986; other versions are due to Hector-Neri Castañeda (1972, 1975, 1977, 1980, 1989a); Terence Parsons (1974, 1975, 1978, 1979ab, 1980; cf. Rapaport 1985a); Richard Routley (1979; cf. Rapaport 1984); and Edward Zalta (1983).)

The relevant point here is that any cognitive agent—human or computer—cannot directly access the contents of the external world. It can only assume that the external world exists, and it can represent it and its contents—together with *non*-existents!—internally.

6.3 METHODOLOGICAL SOLIPSISM AND THE THEORY OF COMPUTATION.

Consider an AI system that has natural-language competence and that gets its input from the external world, i.e., from a user. The point of methodological solipsism is that we could *simulate* this by building in the input (assuming a finite input).

Indeed, this can be done for *any* partial recursive function, according to the Substitution Property, **[???]** which says:

$$(\exists \text{ recursive function } s)(\forall \text{ natural numbers } xyz)[\varphi_{s(x,y)}(z) = \varphi_x(y,z)]$$

Here's what this means: Enumerate the partial recursive functions in some way. Let φ_x be the xth partial recursive function in that enumeration, and suppose that its input is y and z. Then there is another partial recursive function $\varphi_{s(x,y)}$ that is input–output equivalent to φ_x (when y is fixed). Moreover, $\varphi_{s(x,y)}$ is a function with y (i.e., φ's input) stored internally as data. In other words, data can be stored effectively in programs; it needn't be input from the external world.[1]

Note that z is still external data. Can z be "stuffed in" too. I see no reason why not; that would be a special case, where, say, z is an irrelevant constant (one that is read in but ignored).

If we understand methodological solipsism as the Substitution Property, we thus have an argument for methodological solipsism from the theory of computation.[2]

6.4 PHANTOM LIMBS.

Consider the phenomenon of phantom limbs—the feelings that some people who have had limbs amputated have "in" their now-non-existent limbs, feelings that are indistinguishable from the feelings they had before the amputation. This, too, I take as evidence that methodological solipsism is methodologically correct.

In some of the most recent work on phantom limbs, the conclusions are that the experience is *generated* by the brain and is *not* merely a signal from nerve endings at the site of the amputation. The latter would be consistent with methodological solipsism, but would not be a different phenomenon from ordinary sensory perception (except in how it was *interpreted*). According to Melzack (1992: 126), **[???]** "The phenomenon of phantom limbs . . . raises doubts . . . that sensations are produced only by stimuli . . . . The brain generates perceptual experience(s?) **[???]** even when no external inputs occur. We do not need a body to feel a body." If this is not (methodological) solipsism, what is? (Moreover, it suggests how Winston's Problem can be overcome even by an entity with a non-human body.)

6.5 SOME PROBLEMS.

Despite the evidence in support of methodological solipsism, Gilbert Harman has raised objections to it, and Jaegwon Kim has offered some puzzles for which it must provide a solution.

6.5.1 Kim's Puzzles.

Kim considers a physically indistinguishable, perfect replica of oneself, and asks "if two organisms have identical physical features, will they be identical in psychological characteristics as well?" (Kim 1982: 51). Let's consider versions of two of his puzzles:

1. Suppose that I am thinking of Paris, and suppose that my replica is in the same brain state that I am in. Is he also thinking of Paris? (Assume that he has never been to Paris; one can assume that he was created 5 minutes ago.) Should the methodological solipsist should say

[1] Thanks to my colleague Jin–yi Cai for discussions on this. **[CHECK WITH HIM TO CONFIRM MY INTERPRETATION.]**

[2] Perhaps the Kleene Recursion Theorem is relevant, too?

'Yes'? Kim disagrees, since my replica lacks the "historical and cognitive relationship that I have with" Paris (p. 57).

2. Suppose that I see a tree, and suppose that my replica shares my current brain state. Does he see a tree? Again, methodologically solipsistically, should we say 'Yes'? Kim says 'No' (p. 58).

   In this case, it seems to me, the only difference between my replica and me is that in his case, there is no "Sein-correlate"—no external object corresponding to his internal mental representation. Viewed *internally*, we are in the same *psychological* state. Indeed, there is no way for *either* of us to tell whether or not we are "really seeing" a tree—i.e., whether or not there's a tree out there that we're seeing. How might *I* tell? Perhaps by going up and touching the tree? But then my physical brain states will change. Hence, so will my replica's. Hence, he, too, will experience touching-a-tree. From an external observer's point of view, a third person could say which of us was "successfully" seeing a tree. But that's irrelevant to our *psychological* states.

In the second case, we both, so to speak, see a tree *de dicto*, not *de re*. What about in the first case? Kim says that thinking of Paris is *de re*: "some historical-cognitive contact with the city" is "essential" (p. 61). In that case, there are two senses of "thinking of Paris"—thinking of it *de re* and *de dicto*. From the external point of view, I *am* thinking (*de re*) of Paris, but my replica is not. But we are both thinking (*de dicto*) of Paris. Arguably, though, even in the *de re* case, we're both thinking of Paris—after all, our internal physical and psychological states are alike, and Paris exists. There is a different causal story to be told as to *why* each of us is thinking of Paris—but that, methodologically solipsistically speaking, is irrelevant.

Kim has a third sort of puzzle: I *remember* walking in Paris during the summer of 1992. Does my replica *remember* that? According to Kim, "internal psychological states ... [are] those ... whose occurrence does not imply anything about the past or future, or anything existing other than the organism or structure to which the states occur" (p. 60). So, remembering isn't "internal"; so, I *do*, but my replica does *not*, *remember* walking in Paris. But surely there is a psychological state that in *me* is remembering but in my replica isn't. We might—following Kim (p. 64)—call it seeming-to-remember. But that's a bad name, I think, for an internal psychological state that, if there is a Sein-correlate, *is* remembering and, if there isn't, is called 'false belief'.

The upshot of all this is that, as Stephen Stich says, "what knowledge adds to belief is psychologically irrelevant" (Stich 1978, cited in Kim, p. 63). Kim calls this version of methodological solipsism "The Explanatory Thesis: Internal psychological states are the only psychological states that psychological theory needs to invoke in explaining human behavior" (p. 59). As Kim sees it, and I agree, psychological states come in pairs: belief and knowledge, seeming-to-remember and remembering, seeming-to-see and seeing, etc. The first item in each pair is a *de dicto* experience, so to speak; the second is *de re*. The fact that a belief is true (hence, that we have a case of knowledge) or that a memory-experience or perceptual experience is veridical is irrelevant to the cognitive agent' internal psychological state—it is irrelevant from the first-person point of view, irrelevant from the standpoint of *processing*.

6.5.2 Harman's Wide Functionalism.

Harman, as we saw in §4.3, is a supporter of conceptual-role semantics but also a non-solipsist: "Allowance must be made for various connections between concepts and the external world" (Harman 1987: 80). Now, I don't deny that there *are* such connections. I claim merely that such connections tell the *speaker* nothing about his or her language or concepts. Such connections, however, *do* tell a *third* person something, but they give no *first*-person information.

In another essay (written at about the same time), Harman says that "Ordinary psychological explanations are not confined to *reports* of inner states and processes. They often *refer* to what people perceive of the world and what changes they make to the world" (Harman 1988: 15, my italics). But *whose* reports are these—are they first-person or third-person reports? We are only interested in the former. And how literally are we to take 'refer'? I would rather say that the reports are "about" (meant neutrally) what cognitive agents *believe that* they perceive or change. To repeat, we are concerned ony with first-person reports, so it is *internal* beliefs that count, not objects in the external world. And if one were concerned with *third*-person reports, it would *still* be that third person's *internal* beliefs and representations that counted. (Recall Bruner's electricity story in §5.3.)

There is a curious passage immediately following the last quotation:

> Although some ordinary explanations refer to sensory input and some refer to motor output—a hallucination, an attempt to move that fails [recall the blocks-world robot of §2.7.1]—even in these cases there is normally implicit reference to a *possible* environment. (Harman 1988: 15, my italics.)

This is curious for two reasons. First, it seems that a negation is missing, for references to hallucinations and failed moves are hardly references to *actual* input or output—they only *seem* to be. Second, the passage suggests that wide functionalism (Harman's term for the non-solipsistic view) is *intensional*: All psychological states are "about" something, but that something need not exist. With that, I can agree (cf. Rapaport 1978). But I fail to see how that is anything *but* methodologically solipsistic. It says precisely that what is *actually* out there is irrelevant; all that counts is what the cognitive agent *believes* to be out there. But that, of course, is internal.

Harman asks us to consider an

> uninterpreted program: there are three possible input states, **A**, **B**, and **C**. **A** leads to output **X** and **C** leads to output **Y**; **B** has no effect. Do you understand what is going on? No. You need to know how this system is functioning. In fact, the system is a thermostatically controlled air conditioner. ... In order to understand this system you need to know the wide functional story. The narrow functional story is insufficient. (Harman 1988: 17.)

The wide story is the input–output encoding:
A: the current temperature $> 72°$ F
B: $68°$ F $\leq$ the current temperature $\leq 72°$ F
C: the current temperature $< 68°$ F
X: if air conditioner is off, then turn it on, else leave it on
Y: if air conditioner is on, then turn it off, else leave it off.

Now, this suggests that, for Harman, wide functionalism is the claim that there must be an input–output encoding, that a program without the encoding is "insufficient for understanding how the system functions" (p. 17). But insufficient for *whom*? Not for the system itself (from the first-person view), for that's all the system has to go on. Insufficient for *us*? Perhaps, but that encoding, then, is just *our* mapping of the *system's* symbols onto *our* concepts. And how do *we* understand our *own* concepts? Just as the system understands *its* own concepts: in the first-person way. It always comes back to that.

In sum, *narrow* functionalism—what I prefer to call methodological solipsism—is all that is needed to understand how a system understands its own behavior. It is all that is needed to *construct* a *computational* cognitive agent. *Wide* functionalism—the rest of the story, so to speak—at best tells *external* observers something more about what is going on. It gives them the *de re* information that the system itself lacks. But the system itself does not need it for its own purposes. And the external observer only has it indirectly and internally, as well: by having an internal mental model of both the system being studied and the context in which it is situated. And both are "seen" from the third-person observer's own first-person point of view.

And, just in case you forgot, the first-person semantic enterprise is one of correspondences among symbols; hence, it is purely syntactic.

Chapter 7

THE NATURE OF IMPLEMENTATION

7.1 IMPLEMENTATION AS SEMANTIC INTERPRETATION: THESIS.

As I said at the very beginning (§1.2.1), mental states and processes are not merely algorithms but *processes*—in the technical sense of an algorithm in execution.[1] What is the ontological status of a computer process so understood? Note that a computer process must be implemented: It is a physical device *behaving* in a certain *way*; the *way* is described (or specified) by the algorithm. Now, the physical device running the process *implements* the algorithm. The thesis I wish to put forward and examine here is this:

An implementation is a semantic interpretation.

But what is it an interpretation of? In the case at hand, a computer process viewed as an implementation is a semantic interpretation of an algorithm; the algorithm plays the syntactic role. In other cases, that which plays the role of syntactic domain to the implementation's role as semantic domain will be different sorts of things. For reasons that will become clear below, I shall use the term *Abstraction* for the syntactic domain, so my thesis is:

An implementation is a semantic interpretation of an Abstraction.

The bulk of this chapter is devoted to explicating these notions and justifying the thesis.

7.2 GOOD OLD-FASHIONED CARTESIAN DUALISM.

Computational cognitive science, or what John Haugeland (1985: 112) has termed "good old-fashioned artificial intelligence", is, I believe, good old-fashioned Cartesian dualism. The view that mental states and processes are (or are expressible as) algorithms that are *implemented in* the physical states and processes of physical devices is (a form of) Cartesian dualism: The mental states and processes and the physical states and processes can be thought of as different "substances"

[1]How can a mental *state* be a *process*? One way is if the mentalstate is implemented as a certain sequence of neuron firings. Cf. the discussion of Damasio's theories, §§2.8.3, 3.2.2.2.1.

that "interact". How might this be?

It should be clear that an algorithm and a computer are different kinds of "substance". If one considers an algorithm as a mathematical abstraction (in the ordinary sense of the term 'abstraction'), then it is an abstract mathematical entity (like numbers, sets, etc.). Alternatively, if one considers an algorithm as a text expressed in some language, then it is, say, ink marks on paper or ASCII characters in a word-processor's file. An algorithm might even be—and indeed ultimately is—"switch settings" (or their electronic counterparts) in a computer. In any case, that is a very different sort of thing from a very physical computer.

How do mind (or algorithm) and brain/body (or computer) "interact"? By the latter being a semantic interpretation—a model—of the former. More precisely, the *processes* of the brain/body/computer are semantic interpretations of (or models of) the mind/algorithm in the sense of semantics as correspondence. But this is just what we call an implementation. So, an implementation is a kind of semantic interpretation.

Note, by the way, that the mind/algorithm is also a semantic interpretation of the brain/body/computer, since, as we saw in Chapter 4, the correspondence goes both ways. How is a mind implemented? Consider a computer program: Ultimately, the program (as text) is implemented as states of a computer (expressed in binary states of certain of its components). Isn't that purely physical? Yes, but it is *also* purely syntactic; hence, *it* can have a semantic interpretation. An abstract data type, for instance, can be thought of as the semantic interpretation of an arrangement of bits in the computer (cf. Tanenbaum & Augenstein 1981: 1, 6, 45; see §2.3). This Janus-faced aspect of the bit arrangements—thought of both as a physical model or implementation of the abstract algorithm and as a syntactic domain interpretable by the algorithm and its data structures—is simply our old friend the model muddle (§2.6).

Now, is this really good old-fashioned Cartesian dualism? Is mind–body interaction really semantic interpretation or implementation? Or might this semantic/implementational view be more like some other theory of the mind?

- It is not parallelism, since there really is a *causal* interaction: The algorithm (better: the process) *causes* the physical device to behave in certain ways.
- So it's also not epiphenomenalism, either. (And the device—or its behavior—can produce changes in the program, as in the case of self-modifying programs, or even in the case of a system competent in natural language whose knowledge base (part of the software) changes with each interaction.)
- Could it be a dual-aspect theory? Perhaps: Certainly, the physical states and processes are one "level of description", and the mental states and processes are another "level of description" *of the same (physical) system*. But talk of levels of description seems to me to be less illuminating than the theory of semantics as correspondence. More to the point, neither "level" is a complete description of the system: The algorithm is not the process, nor can one infer from the algorithm what the future behavior of the process will be (i.e., the process can behave in ways not predictable by the programmer (cf. Fetzer 1988, 1991). And even a complete physical description of the system would not tell us *what* it is doing; this is one of the lessons of functionalism.

So dualism is at least plausible. Do the physical states and processes produce mental ones?

Here is where the problem of qualia—i.e., subjective qualitative experiences, including pain and physical sensations—enters. We shall have something to say about it later (§7.6.2).

7.3 IMPLEMENTATION AS SEMANTIC INTERPRETATION: EVIDENCE.

> ... the terms of art employed in computer science—... *implementation* ...—will ultimately be definable only with reference to ... attributed semantics" (Smith 1982b: 9.)

Let's briefly review the data we first looked at in §2.3. Some of those pairs of syntactic and semantic domains were clear examples of implementations; others can be so thought of. There are three paradigmatic cases:

semantic domain		syntactic domain
17. a computer program	is an implementation of	specifications
18. a computational process	is an implementation of	a computer program or algorithm
an implementation	is an implementation of	an abstract data type

Thus, we implement a program when we compile and execute it, we implement a set of specifications when we write a computer program, and we implement an abstract data type such as a stack when we write code (in some computer programming language) that specifies *how* the various stack operations (such as *push* and *pop*) will work.

There are several other cases that, while we don't, normally, use the term 'implementation' in discussing them, are clearly of the same type as the paradigms above, e.g.:

semantic domain		syntactic domain
4, 5. a performance	is an implementation of	a musical score or play-script
10. a house	is an implementation of	a blueprint
15. a SNePS node	is an implementation of	a concept

Finally, there are a couple of examples that can clearly be thought of in the same way:

semantic domain		syntactic domain
2. a set-theoretic model	is an implementation of	a formal theory
16. a Sein-correlate (actual object)	is an implementation of	a Meinongian object

My thesis is not only that all implementations are semantic interpretations of a syntactic domain. It is also that all semantic interpretations can be seen as implementations.

7.4 WHAT EXACTLY *IS* AN IMPLEMENTATION?

> On the one hand, we have a very elegant set of mathematical results ranging from Turing's theorem to Church's thesis to recursive function theory. On the other hand,

> we have an impressive set of electronic devices which we use every day. Since we have such advanced mathematics and such good electronics, we assume that somehow somebody must have done the basic philosophical work of connecting the mathematics to the electronics. But as far as I can tell that is not the case. On the contrary, we are in a peculiar situation where there is little theoretical agreement ... on such absolutely fundamental questions as, What exactly is a digital computer? What exactly is a symbol? What exactly is a computational process? Under what physical conditions exactly are two systems implementing the same program? (Searle 1990: 24.)

7.4.1 Implementation in Computer Science.

Let's look first at the notion of implementation in its home territory: computer science. This is not to say, I hasten to add, that the term is not used elsewhere or that it does not antedate computer science—which, as we will see, it certainly does. But since the term is so ubiquitous in computer science, it is a good place to start. Once we have a clear idea of how computer scientists use the term, we'll look at how ordinary folks use it.

Given the ubiquity, it is rather surprising how few texts even *try* to define the notion. For instance, all that Michael Marcotty and Henry Ledgard say in *Programming Landscape* (1986), a standard text on programming languages, is that "the **realization** of a programming language in a computer system is called the *implementation* (Marcotty & Ledgard 1986: 8, my boldface). 'Realization', of course, is left undefined. Taken literally, it means "to make real", where 'real' is opposed to 'imaginary' or perhaps 'abstract'. Before exploring this a bit further, note, first, that this makes it seem that the physical medium is important and, second, that to "realize" could be to establish a Sein-correlate, which can be generalized to different possible worlds. Thus, intelligence, say, could be realized in several different (physical) media: This is the notion of "multiple realizability", to which we shall also return.

"Realization" itself is a rather interesting notion. According to the new *Oxford English Dictionary*, 'real' comes from the Latin for "pertaining to things", and its philosophical meaning, in part, is "having an existence in fact and not merely in appearance, thought, or language" (Simpson & Weiner 1989, Vol. **00**, **[???]** p. 272). What is it that is made real when it is "realized"? Presumably, something that exists "merely in appearance, thought, or language"—in short, something "Abstract". To *realize* is, in part, "To make real, to give reality to (something merely imagined, planned, etc.) ... In common use from *c* 1750 with a variety of objects, as ideas or ideals, schemes, theories, hopes, fears, etc. ..." (Simpson & Weiner 1989, Vol. **00**, **[???]** p. 277). Note how *psychological* or *intentional* these realizable things are. Note, too, especially in connection with the example of a performance of a play, that where English talks of a *director*, French talks of a *réalisateur* (a realizer): At least for francophones, plays and movies are *implementations* (of scripts).

7.4.1.1 Hayes's notion of implementation.

One computer-science text that says a bit more about implementation is John P. Hayes's *Computer Architecture and Organization* (1988). The term first occurs in the following passage:

> With the advent of the [IBM] System/360, the distinction between a computer's architecture and its **implementation** became apparent. As defined by the System/360 designers . . . , the *architecture* of a computer is its **structure and behavior as seen by an assembly-language programmer** . . . . The *implementation* . . . refers to the **logical and physical design techniques** used to **realize** the architecture in any specific instance. Thus all the members of the S/360–370 series share a common architecture, but they have many different implementations. For example, some S/360–370 CPUs employ fast hardwired control units, whereas others use a slower but more flexible microprogrammed approach to implementing the common instruction set. (Hayes 1988: 47, my boldface.)

Architecture is concerned with structure and behavior; these are functional, Abstract aspects. This is not to say that it is not *detailed*, however, since the architecture is "seen by an assembly-language programmer," who must know all about the details of registers, control, etc., although he or she does not have to worry about what a register looks like or how the control is actually carried out.

Implementation is concerned with "logical and physical design techniques". I am not sure what 'logical' means here, but 'physical' is clear: Implementations are the physical realizations of Abstractions. I would call a *physical* realization a special case, however. The full explication of 'implementation' requires a third term besides the implementation and the Abstraction:

I is an implementation of A in medium M

where A is the Abstraction and M could be physical or set-theoretical, etc. For instance—as we will see below—it is common in the study of data structures in computer science to talk about implementing a stack by means of a linked list, implementing the list in a programming language (say, Pascal), "implementing"—i.e., compiling—the Pascal program in some machine language, and then implementing the machine-language program in a real computer. As we progress along this sequence (this correspondence continuum), the implementing media begin as Abstractions themselves and gradually take on a more "physical" nature.

Perhaps you will object that program compilation should not be treated as an implementation. Hayes, however, would not object: "a sequence of . . . [machine] instructions is needed to *implement* a statement in a high-level programming language such as FORTRAN or Pascal" (Hayes 1988: 209, my italics). So, to implement is to "realize" *in* some medium, which might be a physical medium or could be some *domain* or *language*. To implement is to *construct* something, out of the materials at hand, that has the properties of the Abstraction; it could also be to *find* a counterpart that has those properties. Both tasks are semantic.

Hayes, indeed, speaks of semantics in this context:

> Because of the complexity of the operations, data types, and syntax of high-level languages, few successful attempts have been made to construct computers whose machine language directly corresponds to a high-level language . . . . There is thus a *semantic gap* between the high-level problem specification and the machine instruction set that **implements** it, a gap that a compiler must bridge. (Hayes 1988: 209, my boldface.)

What is the gap? Presumably, that (say) the specific operations, data types, etc., of the high-level language don't correspond directly to anything in the machine language: Pascal, e.g., has the

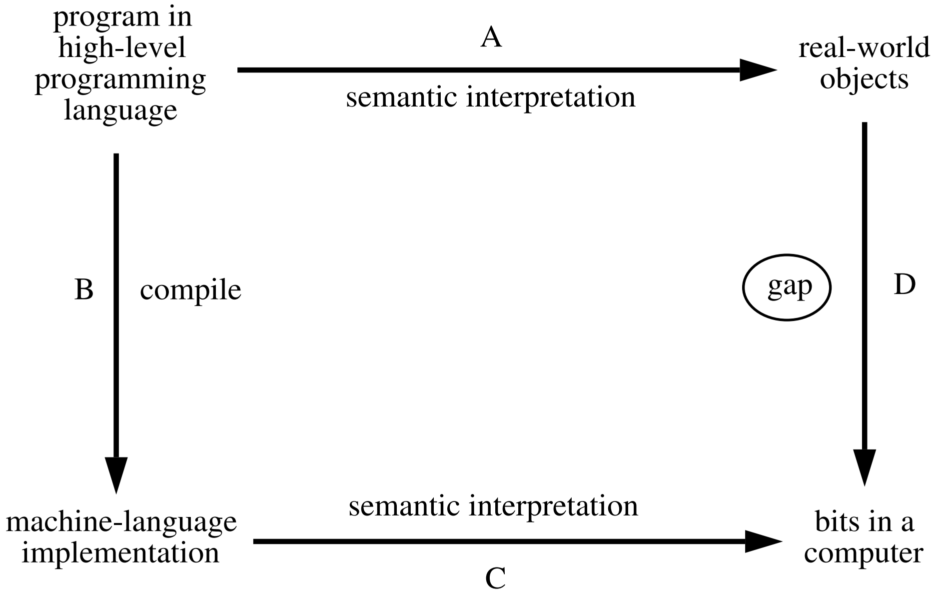


Figure 7.1: A program in a high-level programming language is semantically interpreted by real-world objects. It is also compiled into a machine-language implementation, which, in turn, is semantically interpreted by bits in a computer. What is the relation between these two semantic interpretations?

"record" data type, but my VAX's machine language probably doesn't. So, a compiler is needed to show how to construct or implement records in the machine language. (This notion of construction is, perhaps, related to the notion of the construction of the rational numbers from the integers and, hence, to the notion of theory *reduction*. We will return to this in §7.5.3.)

Why does Hayes call this a *semantic* gap? It's a bit like the fact that one natural language might not have a single word corresponding to some single word in another natural language. John Sowa (personal communication, 29 November 1993) gives the example of Russian, which has a term, 'ruka', referring to what in English has to be referred to as the hand+forearm. Of course, one can translate between the languages by defining the word in terms of others (perhaps with a cultural gloss (cf. Jennings 1985, Rapaport 1988: 102). But why is this *semantic* rather than *syntactic*?

A possible interpretation[2] is shown in Figure 7.1, which specifies four relations:

A. Presumably the semantic interpretation of a program written in a high-level programming language is the relation between, on the one hand, data structures (say) in a Pascal program (say, a record representing a student viewed as consisting of a name, an age, a class, a major, a student-number, and a grade-point average) and, on the other, an actual student in the real world.

B. The compilation relation is, or includes, the relation between that student-record data structure and a construct of data types in the machine language. Note that both A and

[2] Due to my colleague Bharadwaj Jayaraman (personal communication).

B, on my theory, are semantic relations.

C. At first sight, it seems odd to semantically interpret the machine-language program by bits. Why not map the machine-language program into the real-world objects? But, after all, if all semantic relations are merely correspondences (and vice versa), then this relation between the machine-language program and bits is just another one. After all, we could also have mapped the Pascal program into computer bits—in fact, via B and C, we have! So, given a machine-language program, one can interpret it in terms of bits in the computer. Arguably, in fact, having these two distinct interpretations of two distinct (albeit input–output—equivalent) programs is appropriate. Where the machine-language program talks of registers, the Pascal program talks of "students" (or student-records). So it is appropriate to understand the Pascal program as a "mathematical model" of such real-world objects as students and to understand the machine-language program as a "mathematical model" of such (also real-world) objects as bits in a computer.

D. So the gap concerns the relation between the real-world objects (such as students) and computer bits, since *both* are semantic interpretations of the Pascal program.

What, then, is this relation D? It could be *simulation*: The computer bits simulate the student. Or it could be *implementation*: The computer bits are a computer implementation (an implementation in the medium of the computer) of the student. Before exploring some of these options further, let's look more closely at one of the standard uses of 'implementation' in computer science: the implementation of an abstract data type.

7.4.1.2 Abstract data types.

The notion of an abstract data type and its "implementation" is one of the most common uses of 'implementation' in cognitive science. There is a relatively informal use of the notion, as it appears in programming languages such as Pascal and as it is taught in introductory computer-science courses, and there is a more formal, mathematically precise use. Let's look at the informal one first, informally, in order to provide a bit of background.

7.4.1.2.1 The informal notion of implementation. A stack is a particular kind of data structure, often thought of as consisting of a set of items structured like a stack of trays in a cafeteria: New items are added to the stack by "pushing" them on "top", and items can be removed from the stack only by "popping" them from the top. Thus, to define a stack, one needs (a) a way of referring to its top and (b) operations for pushing new items onto the top and for popping items off the top. That, more or less (mostly less, since this is intended to be informal), is a stack defined as an abstract data type.

Now, Pascal does not have the stack as one of its built-in data types (as it does arrays, records, or sets). So, if you want to write a Pascal program that manipulates data structured as a stack, you need to "implement" a stack in Pascal. This can be done in several ways; i.e., there are several ways to implement the abstract data type *stack* in Pascal. Here's one way:

1. A stack, s, is implemented as a 1-dimensional array, $A[0], \ldots, A[n]$, say, for some n;

2. $top(s)$ is defined to be a 1-argument function that takes as input the stack s and returns as output $A[0]$ (i.e., $A[0]$ is the implementation of the "top");

3. $push(s,i)$ is defined to be a 2-parameter procedure that takes as input the stack s and an item i (of the type allowed to be in the array), and yields as output the stack modified so that $A[0] := i$, and $A[j] := A[j-1]$ (i.e., each item on the stack is "pushed down"); and,

4. almost finally, $pop(s)$ is defined to be a 1-argument function that takes as input a stack s and returns as output the item on the top of s (i.e., $top(s)$) while moving all the rest of the items "up" (i.e., $A[j] := A[j+1]$).

I said "almost finally" because—as should be obvious—there is some bookkeeping that has to be taken care of:

5. We have to specify what happens if the stack "overflows" (as when we try to push an $(n+2)$nd item onto a stack implemented as an $(n+1)$-element array).

6. We have to specify what happens to the "last" item when the top is popped (does the array cell that contained that item still contain it, or does it become empty?), etc.

These (as well as the limitations due to the type allowed to be in the array) are what are called "implementation details", since the abstract data type *stack* "doesn't care" about them (i.e., doesn't—or doesn't *have to*—specify what to do in these cases).

Here's another way to implement a stack in Pascal. Do everything as before, but let $top(s) := A[n]$. This implementation of the *stack* abstract data type is "inverted" with respect to the first one. The inversion, however, is (a) a ("mere") implementation detail and (b) undetectable in the program's input–output behavior. (Start thinking about inverted spectra here; we'll come back to them in §7.6.2.)

And here's a third way: Use Pascal's *pointer* data type to implement a stack as a "linked list". I won't bore you with all the details, but here are a few. First a linked list ('list', for short) is itself an abstract data type. It is a sequence of items whose three basic operations are (1) $first(l)$, which returns the first element on the list, (2) $rest(l)$, which returns a list consisting of all the original items except the first, and (3) $make\text{-}list(i,l)$ (or $cons(i,l)$), which recursively creates a list by putting item i at the beginning of list l. Lists can be implemented in Pascal by, e.g., 2-dimensional arrays (here, the first item in each two-cell unit of the array is the list item itself, and the second item in the two-cell unit is an index to the location of the next item) or by means of "pointers" (each item on the list is implemented as a two-element "record", the first element of which is the list-item itself and the second element of which is a pointer to the next item). A stack, s, finally, can be implemented as a list, l, where $top(s) := first(l)$, $push(s,i) := make\text{-}list(l,i)$, and $pop(s)$ returns $top(s)$ and redefines the list to be $rest(l)$.

So: stacks can be implemented as arrays or as lists, and lists can be implemented as arrays or as records with pointers. Abstract data types can implement other abstract data types, or they can be implemented "directly" in the given data structures of a programming language. What's going on here?

7.4.1.2.2 The formal notion of implementation. To see what's going on, we need to look at some of the more formal approaches to the definition and implementation of abstract data types.

7.4.1.2.2.1 ***Guttag, Horowitz, and Musser.*** In "The Design of Data Type Specifications" (1978), John V. Guttag, Ellis Horowitz, and David R. Musser assert that "the process of design (of data types) consists of specifying ... operations to increasingly greater levels of detail until an executable implementation is achieved" (p. 61). So, the implementation appears to be merely a more detailed version of the original "specification". The implementation details are essential for *executability* but not, presumably, for specifiability. So the implementation details serve a purpose, but one distinct from the original abstraction.

What is the abstraction? "A *data type specification* (or abstract data type) is a representation-independent formal definition of each operation of a data type. Thus, the complete design of a single data type would proceed by first giving its specification, followed by an (efficient) implementation that agrees with the specification" (p. 61). So, implementations—the *detailed* specification of the operations—must "agree with" the *abstract*—or *undetailed*—specification; the implementation must *satisfy* the definitions. That, of course, needs to be made precise, but it is more than suggestive of semantic interpretation. If the abstraction is supposed to be "representation-independent", then perhaps the implementation *is* the (or, a) representation. (Note that there can be more than one implementation; at the very least, there can be "efficient" and inefficient ones.)

Guttag et al. give "a brief example of the implementation of one data type, Queue ..., in terms of another, CircularLists" (p. 74). So, as we noted before, abstract data types can implement each other. This is done as follows: "We first give, in a notation very similar to that for the specification, an implementation of the Queue type consisting of a *representation* declaration and a *program* for each of the Queue operations in terms of the representation" (p. 74). In the example, the representation "medium" is CircularList, and the "programming language" consists of the operations of CircularLists. So, an implementation of an abstract data type consists of a representation and programs, where the programs implement the abstract data type's operations. This is done as follows: Each operation of the abstract data type is ... defined? explicated? implemented? ... in terms of an operation of the implementing medium (the implementing abstract data type), *after* first representing each abstract-data-type entity (term) by a term of the implementing abstract data type. So, terms get interpreted by, or mapped into, elements of the interpreting domain, and predicates (operations) are mapped into predicates of the interpreting domain. So, implementation *is* semantic interpretation.

7.4.1.2.2.2 ***Goguen, Thatcher, and Wagner.*** A somewhat more detailed and philosophically sophisticated approach is to be found in J. A. Goguen, J. W. Thatcher, & E. G. Wagner's "An Initial Algebra Approach to the Specification, Correctness, and Implementation of Abstract Data Types" (1978). The mathematical details they present are, I think, irrelevant to our inquiry in this chapter, but the overall picture they offer is useful, so let me attempt to summarize it here.

They begin by observing that "the term *abstraction* in computer science ... has been used in at least three ways which are distinct but related" (p. 82):

1. An abstraction is "a mathematical model or description of something" (p. 82).
2. An abstraction is "the process (or result) of generalizing" (p. 83).
3. An abstraction is "a concept" considered "independent[ly] of its representation" (p. 83).

Examples of (1) are "'abstract machines' as opposed to real hardware" (p. 82) and "abstract implementations", as "when one uses sets, sequences, or other mathematical entities to model some computational process or structure" (p. 83). So, a mathematical model is an abstraction of some real-world entity; as such, the abstraction seems to play the syntactic role. On the other hand, the implementation of a queue by a circular list (or a stack by a linked list) is an "abstract implementation", yet here it clearly plays the *semantic* role. This, I take it, is a further example of the muddle of the model in the middle.

In sense (2), abstractions ignore details. This contrasts nicely with the notion of an implementation as *providing* details. Presumably, an abstraction in sense (1) might ignore details, and hence be an abstraction in sense (2), but not necessarily. For instance, although the (admittedly informal) abstract notion of stack that I presented in §7.4.1.2 ignored the details imposed by the finiteness of the stack, we could have had an equally abstract presentation that payed attention to those details (yet could be implemented as a (finite) array, an "inverted" array, or a (finite) list). However, *merely* ignoring details does not by itself yield an abstract model in sense (1), because it might not be a *mathematical* model.

It is the third sense of 'abstraction' that Goguen et al. take to be the relevant one for abstract data types (cf. p. 81). The "representation" that such an abstraction is independent of has to do with notation, or the manner in which it is expressed:

> For example, "abstract syntax" considers syntactic structure independently of whether it is represented by derivation trees, parenthesized expressions, ... or whatever. This notion of abstract syntax is useful ... in specifying the semantics of a programming language in a manner independent of how it is implemented ....
>
> More to the point, an abstract data type is supposed to be independent of its representation, in the sense that details of how it is implemented are to be actually hidden or "shielded" from the user: He [sic] is provided with certain operations, and he only needs to know what they are supposed to do, not how they do it. (Goguen et al. 1978: 83; cf. Parnas 1972.)

That is, the programmer can deal directly with the abstract data type and ignore its implementation; one deals with it at a "high level". Consistent with our view that an implementation is a semantic interpretation, Goguen et al. observe: "Note that what is usually called an 'abstract implementation,' that is, an implementation described by sets, sequences, etc., is *not* an 'abstraction' in the above sense; rather, it is a *particular*, but rather undetailed, implementation" (p. 83). So, an abstraction in sense (1) is not necessarily an abstraction in sense (3). It is undetailed, presumably because the implementing medium (the implementing abstract data type) is *itself* abstract (in sense (2)). Still, the mathematical model is a semantic model.

Now, what is this abstraction of the third kind? Goguen et al. note that it has to do with equivalence classes, or what they call "isomorphism classes" (p. 83). They define an abstract data type as "the isomorphism class of an initial algebra in a category" of many-sorted algebras (pp. 88, 90). And they note that "An implementation is necessarily made within a specific framework, such as a particular programming language or machine" (p. 135); i.e., an implementation requires an implementing *medium*, or "framework".

Their mathematical "approach is to model an implementation framework as an algebra, with the elements of the carrier(s) being concrete data representations (machine states, primitive data types) and its operations the given basic operations (machine operations, basic instructions, programs) in these data representations" (p. 135). Note that they are *modeling* the implementing *medium* and that they do so by the *same* kind of entity as for an abstract data type, namely, an algebra! The implementation *itself*, of course, is something "physical"; it is merely being *described* algebraically.

The heart of the matter is expressed by them in their mathematical set-up as follows:

> Let B denote the implementation algebra .... [Let $T_{\Sigma,\epsilon}$ be] the specification algebra. The question now is, What relationship between $T_{\Sigma,\epsilon}$ and B constitutes an *implementation*? (p. 136.)

This is indeed the question: What is the relationship between an Abstraction (a "specification algebra") and an Implementation (an "implementation algebra")? (Note, as with Smith, that B itself is (merely) a representation or model of the actual, physical implementations.) Goguen et al.'s answer is that the relationship is a structure consisting of B, a mapping from (roughly) $T_{\Sigma,\epsilon}$ to B, and a "congruence" (a family of equivalence relations on (roughly) T's image in B) (p. 138). The core of this is, first, the mapping from the Abstraction to the Implementation, which is, on my theory and consistent with the view of Guttag et al., a *semantic interpretation*, and, second, the "congruence". The latter is a very special, intricate kind of isomorphism, one that "factors out" (or "divides out"—they use quotient spaces) the "implementation details". So, B (or that which B is a mathematical model of) implements an Abstraction T if and only if B is a domain of semantic interpretation of T, ignoring the implementation details.

An example of what I think is going on might help. Consider the abstract data type *stack*, and consider two specific implementations of it in Pascal, using an array $A[0], \ldots, A[n]$ with *top* implemented in one as $A[0]$ and in the other as $A[n]$. In both implementations, *top* is implemented as a specific element of the array. That it is $A[0]$ in one and $A[n]$ in the other is an implementation detail.

But can those details really be ignored? In this case, perhaps; in others, perhaps not—we'll return to this (§7.6). Before we do, though, we need to broaden our scope and consider to what extent the notion of an implementation is applicable *outside* of computer science and then to look at other candidate interpretations of 'implementation'.

7.4.2 Implementation Outside of Computer Science.

7.4.2.1 Music.

Some of the clearest examples outside of cognitive science of what could be called 'implementation' come from music (cf. §2.3, example 5). This ought not to be surprising: After all, a music score is very much like a computer program or algorithm, and the musician-plus-instrument (or conductor-plus-orchestra) plays a role very much like that of the computer. A musical score is *not*, of course, *mathematically* an algorithm, since much is left open to "interpretation" by the musician (e.g, tempo, dynamics, optional repeats, phrasing, etc.). Nonetheless, it *is* a set of "instructions" which, when followed or executed, produce a certain output. The "process" consisting of the musician playing that music on an instrument can plausibly be said to *implement* the score. The score can be thought of (indeed, it *is*) a piece of syntax; the playing of the score provides a "semantic interpretation" of it.

Now, as we saw above (§7.4.1.1), an implementation requires an implementing medium. And, as should be evident, there can be many different media, hence many different implementations (the common core of which can be captured by the mathematical techniques of Goguen et al.). We find the same thing in music: A given score can normally be played on a variety of instruments, modulo a few changes necessitated by the nature of the instrument. Such changes, as well as the particular features of the instrument, constitute "implementation details". Often, these change the nature of the work, for good or bad: "a [piano] transcription [of a symphony] can hold a prism up to a familiar work, showing it in a new light" (Pincus 1990: **00**). **[???]** That is, a piano transcription of a symphony is an *interpretation* of it—or, rather, *another* interpretation of "the work", i.e., of an abstract data type (the score) of which both the symphony *and* the piano transcription are (semantic) interpretations or implementations. The implementation is also, of course, an "interpretation" in the ordinary sense: In an essay on "historically accurate performances" of music, Charles Rosen (1991) speaks of "the essential gap between the composer's conception of a work of music and the *multiple possibilities of realizing it in sound*" (p. 50, my italics). The "conception" is the abstract data type; the "multiple possibilities" are different implementations. Much the same can be said, *mutatis mutandis* for scripts and productions of plays (or scripts and movies) (again, recall §2.3, examples 6, 7, 10).

Music, though, has another aspect to it reminiscent of the situation with computers. On a radio quiz show called "My Music" (radio station CJRT FM 91.1, Toronto; 28 November 1990), a tune was played on a piano for the panel to identify. One panelist "misidentified" the piece. The host said, "You were listening to a tune from Puccini's *Girl of the Golden West* but thought you were listening to Lloyd Weber's *Phantom of the Opera*," to which the panelist drily replied, "It was both." Another, perhaps more familiar, example would be "76 Trombones" and "Good Night My Someone", both from *The Music Man*, both the same tune, albeit different words, tempi, and arrangements—in short, two different implementations of a single Abstraction.

Here, though, we can see another phenomenon at work, too, for these cases also seem very much like that of the Morning Star and the Evening Star, or the conversation that is simultaneously about mathematical lattice theory and chemical lattices (§2.7.1), or the computer program that can be taken either (or: both) as playing chess or (or: and) analyzing a Civil War battle. Given the Morning Star/Evening Star way of thinking about it, it would follow that semantic interpretations and, hence, implementations, are (perhaps not surprisingly) intensional entities and that we should expect to find *de re*/*de dicto* phenomena. For instance, the panelist on *My Music* was listening to (or perceiving) *de re* an Abstraction; *de dicto*, however, he was listening to Puccini, not Lloyd Weber. *De dicto*, one describes oneself as perceiving an interpretation. One interprets what one perceives; possibly, one *can't just* perceive the Abstraction *de re*.

7.4.2.2 Language.

Another non–computer-science example of implementation was discussed by Sellars:

> [In the context of chess,] . . . attention must be called to the differences between 'bishop' and 'piece of wood of such and such shape'. . . . [the former] belongs to the rule language of chess. And clearly the ability to respond to an object of a certain size and shape *as a bishop* presupposes the ability to respond to it as an object of that size and shape. But it should not be inferred that 'bishop' is 'shorthand' for 'wood of such and such size and shape' . . . 'Bishop' is a counter in the rule language game and participates in linguistic moves in which . . . the . . . longer expression does not . . . . (Sellars 1955/1963: 343, §56.)

"Being a bishop" is a nice example of what I am calling an Abstraction. Here, a bishop is implemented as a certain piece of wood (cf. also the example in §3.2.1). It could also, as Sellars observes, be implemented by a Pontiac if the chess game is played in Texas, where everything is supposed to be bigger:

> . . . the term 'bishop' as it occurs in the language of both Texas [where it is "syntactically related . . . to expressions mentioning different kinds of cars" (p. 344, §59)] and ordinary chess can be correctly said to have a common meaning—indeed to mean the bishop role, embodied in the one case by pieces of wood, and in the othe by, say, Pontiacs . . . . (Sellars 1955/1963: 348, §62.)

This situation is depicted in Figure 7.2. Here, we have an Abstraction (Chess) and two implementations (the ordinary Staunton pieces and the Texas pieces). We assume that the pieces that play the role of the bishop are both *called* 'bishops'; 'bishop' means the same thing in both implementations, namely, the Bishop Abstraction. That role is "embodied as"—i.e., *is implemented by*—a Pontiac in Texas and a certain ♗-shaped piece of wood in the Staunton set. The words 'bishop' as they occur in the two different languages refer to different entities (the language-entry and -departure rules in Sellars's language games differ).

Language provides non–computer-science examples of implementation in a variety of ways. For one thing, as Roy Harris has observed (Harris 1987: xi, Ch. 6), words can be considered as representations—hence, implementations—of ideas. For another, if language can be thought of as an Abstraction (as, perhaps, Chomsky's theory of universal grammar would have it), then it can be implemented in a variety of ways: first, by spoken languages (implemented in the medium of speech) as well as by signed languages (implemented in the medium of space) (cf. Coughlin 1991, cited in §2.3, example 18), and, second, in many ways in both spoken and signed languages (e.g., French, English, etc., and American Sign Language, British Sign Language, etc.).

Another example from language is also due to Sellars. At the beginning of "Some Reflections on Language Games" (1955/1963), he distinguishes between "*obeying* rules[, which] involves using the language in which the rules are formulated, [and] *conforming* to rules[, which] does not" (p. 322, §4). He goes on in the same passage to observe that "once one has learned" the metalanguage in which the rules are formulated, "one may come to *obey* the rules for" the language whose rules for use are formulated in the metalanguage. This is precisely how I view the "classical" or "symbolic" paradigm in AI in general and in natural-language competence in particular. It is a feature of the "multiple realizability" of natural-language competence: In us, *conceivably*, natural-language competence is implemented by rule-*conforming*; in a computer, it could be implemented by rule-*following*. This assumes, of course, that there are rules to which we conform. But, at least, the

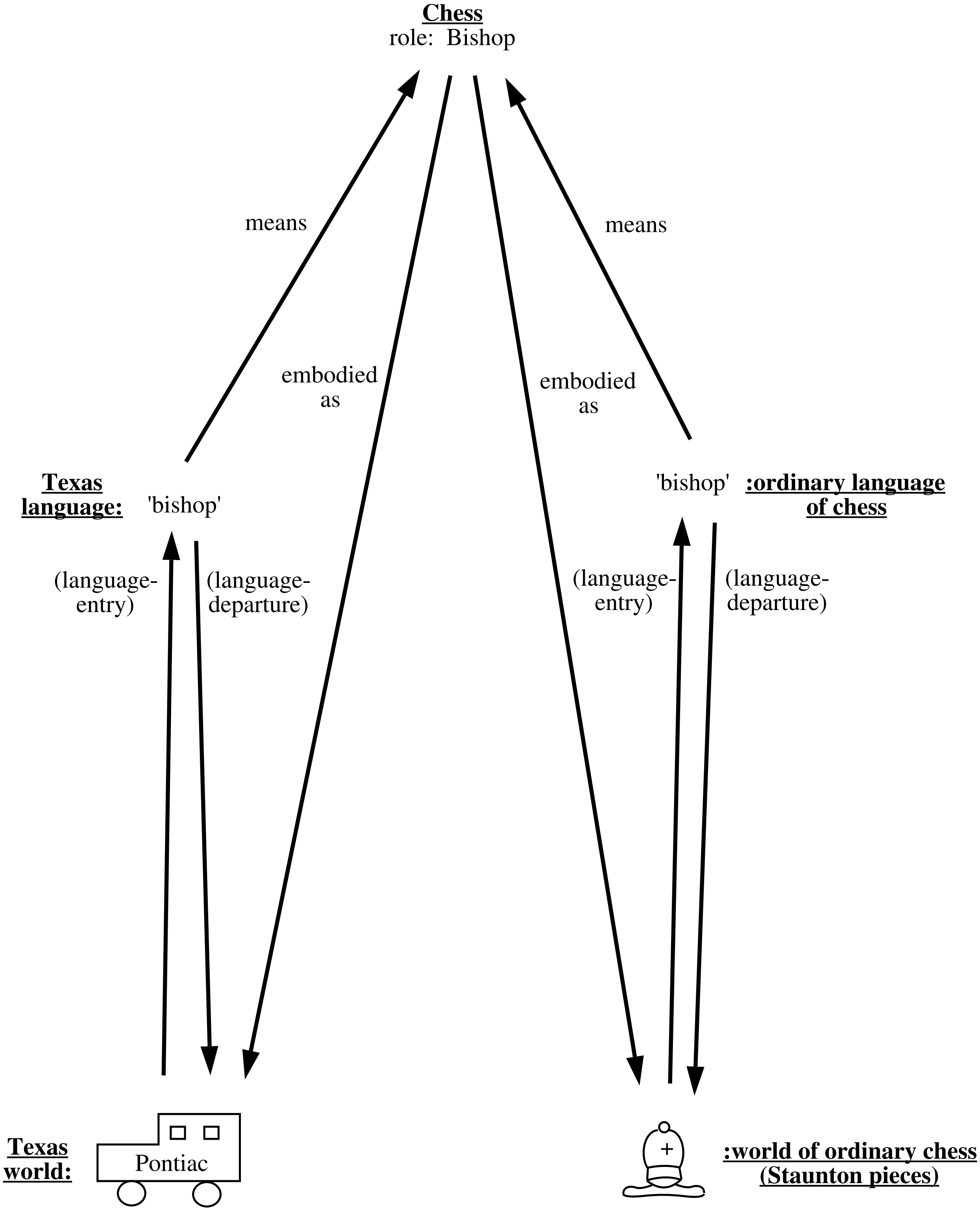


Figure 7.2: The Chess Abstraction and two of its implementations (cf. Sellars 1955/1963: 346).

search for such rules is a worthwhile scientific goal.[3]

Finally, there is the issue of the medium. I once saw a discarded library catalog card and thought, "How sad," because that's where the information "really" is, not in the electronic "card catalog", which is just a bunch of 0s and 1s (and not really even that). But that's wrong, of course. The information—the very *same, identical* information—is stored (*implemented*) electronically (but cf. N. Baker 1994). After all, the information on the card is just a bunch of ink marks on paper (cf. the "mystery of reading", §5.1). The medium *isn't* the message; rather, the message is *implemented* in the medium. There is a sense, though, in which the medium *is* at least *part* of the message. That's the sense in which qualia reside in the implementation details (again, cf. N. Baker 1994), or perhaps the sense in which a robot with a different body might have different concepts from humans. (On the former, see §7.6.3, below; on the latter, see §3.2.2.2.4, above.)

7.4.2.3 Mind.

My last example of implementation outside of computer science is the mind–body problem. It would be too simplistic to say that the mind is implemented in the brain, but that is certainly the central insight. If AI "succeeds", then in its Golden Age we will have a set of algorithms for cognition that, when implemented, yield a cognizing entity. If the algorithms are strongly equivalent to those used by the human mind (i.e., not merely input–output equivalent, but equivalent in detail), then we would say that they were implemented in the human brain:

> Symbolists emphasize that the symbolic level (for them, the mental level) is a natural functional level of its own, with ruleful regularities that are independent of their spec. **[???]** physical realizations. For symbolists, this implementation independence is the critical difference between cognitive phenomena and ordinary physical phenomena .... (Harnad 1990: 336.)

So, implementation independence might be taken as a mark of the mental. There are two ways of looking at this: (1) *Being syntactic* is a mark of the mental. (2) *Being interpretable* is a mark of the mental. The second of these subsumes Brentano's claim that intentionality is the mark of the mental.

7.4.3 Definitions.

Before summing up, let's take another look at the *Oxford English Dictionary.* The *noun* 'implement' comes from the Latin for "a filling up", as in "that which serves to fill up or stock (a house, etc.)", and from the Old French for "to fill, fill up" in the sense of "completing" (Simpson & Weiner 1989, Vol. VII, p. 721). This suggests "filling in the details", which an implementation in the sense we are concerned with certainly does.

The *verb* is of more recent origin and has three senses, all with citations beginning in the 19th century (p. 722):

[3]Unfortunately, Sellars goes on to argue that the obeying–conforming dichotomy is false (p. 325, §12), but he replaces it with a distinction between "'pattern governed' and 'rule obeying' behavior" (p. 327, §16; cf. §§18–19), which will do as well.

(a) "To complete, perform, carry into effect (a contract, agreement, etc.); to fulfil (an engagement or promise)." This is the earliest sense to be cited (1806)—implementing an obligation.

(b) "To carry out, execute (a piece of work)." Here, the citation is from 1837: implementing an invention.

(c) "To fulfil, satisfy (a condition)." This was used as early as 1857: implementing the "mechanical requisites of the barometer ... in ... an instrument".

Senses (b) and (c) seem closest to our concerns: Sense (b) relates to an Abstraction, and (c) relates to the implementation of an Abstraction—to satisfying the conditions of the Abstraction, or having the properties of the Abstraction. More recent senses (with citations from 1926 and 1944) don't clarify much; curiously, none of the citations come from computer science.

Recall that "implementation" is a relational notion, whose full context is always:

I is an "implementation" of an "abstraction" A in some "medium" M.

I have suggested that the notion of *Abstraction* be a generalization of the notion of an abstract data type, and we have seen from examples that syntactic systems and conceptual-role systems are Abstractions ripe for "implementing". Must Abstractions be abstract, that is, non-spatiotemporal? If so, then they would contrast nicely with a *physical* or *concrete* interpretation of an "implementation"—i.e., with the "medium" always being spatiotemporal. But we have seen that one Abstraction can implement another. So this characterization won't do. Instead, I suggest that we leave the notion unrefined for now except as that which can be implemented in some medium.

What about the medium? It could be abstract or concrete, giving rise to two varieties of implementation. An "abstract implementation" would be a specification, a filling-in of details, of an Abstraction. For instance, in top-down design, each level (except possibly for the last) is an "abstract implementation" of the previous one: I begin preparing my courses with a bare-bones course outline and successively refine it by adding details; or: I start solving a problem algorithmically by writing an algorithm in "pseudocode" and, by "stepwise refinement", fill in the details (e.g., pseudocode the procedures), until I finally encode it in, say, Lisp.

A "concrete implementation" would exist in a physical (or spatiotemporal) medium. It would necessarily have more details filled in, namely, those due to, i.e., contributed by, the medium. For example, my actually standing in front of the class, lecturing, is a concrete implementation of my final course outline. The actual words I say, the actual piece of chalk I use, etc., are all implementation details, filled in to "the last detail" by the very nature of the real, spatiotemporal events. Similarly, the actual execution of my Lisp program (perhaps after having been compiled into—i.e., further implemented in—machine language)—the *process*—is its concrete implementation. Note that both abstract and concrete implementations are semantic interpretations.

But this is only the tip of the iceberg. Is "implementation" a concept *sui generis*? Or is it just another name for something more familiar, such as "instance", "exemplification", "reduction", "supervenience", etc.? Let us consider some of these.

7.5 POSSIBLE INTERPRETATIONS OF "IMPLEMENTATION".

The problem in trying to determine whether the notion of "implementation" should be assimilated to some other, perhaps more familiar, notion is that there is very little agreement over the proper characterization of those other, candidate notions, or even over terminology. For instance, it seems clear that an implementation of an Abstraction is *not* an "instance" or "instantiation" of the Abstraction, because two Abstractions (e.g., two abstract data types) can implement *each other*. As we saw, the abstract data type Record can be used to implement the abstract data type List. Moreover, the abstract data type List can be used to implement the abstract data type Record. And, though there is probably no good reason to do so, one could, perversely, implement lists by records that are themselves implemented by lists. And so on. Yet "instantiation" is normally thought of as an asymmetric relation. In spite of this, we find a recognized authority on implementations, Guttag et al., saying that an implementation *is* an "instance" (Guttag et al. 1978: 62). In this chapter, I do not expect to be able to resolve the issues, only to set them out.

7.5.1 Implementation as Individuation.

Let's begin (following Castañeda 1975b) by distinguishing between "individuation" and "differentiation". Recall that early ancestor of semantic networks, Porphyry's Tree, in which a universal, such as a genus, is analyzed into sub-genera or species by means of a "specific difference" or *differentia*. Thus, for example, the *differentia* Rational applied to the genus Animal yields the species Human (= Rational Animal); all other, non-human, animals are *not* Rational. Thus, Humans are *differentiated* from non-Humans. As a category, Human is "lower" than Animal; it is more "specific"—it has an extra defining property, namely, being rational. Human is itself a universal—as it happens, an *infimum species*, i.e., a category that is not analyzed into subcategories but into concrete *individuals*, e.g., Plato, Sappho, you, me.

What is the analogue of a *differentia* that, when applied to an *infimum species* yields an individual? John Duns Scotus called it 'haecceity', or "thisness". "Instantiation" is the relation between any level of Porphyry's Tree and the level *below* it; "differentiation" is a relation between subcategories (or members) of a single category. Thus, just as Human is differentiated from non-Rational Animal, so Plato is differentiated from Sappho and you from me. And just as Human is instantiated from (or, is an instance of) Animal, so Plato, Sappho, you, and I are instantiated from (or, are instances of) Human. And Plato et al., unlike Human et al., are "individuals": "Individuation" is the relation between an *infimum species* and its individuals. As Gracia puts it, "I regard something as an individual if and only if it is a noninstantiable instance of an instantiable, while I regard universals as capable of instantiation, that is, as instantiables" (Gracia 1990: 503).

Thus, perhaps implementations are individuals, and Abstractions are universals. That does seem to hold for *concrete* implementations. But it fails to hold for *abstract* implementations, and it only works when there is a hierarchy or linear ordering of successively more detailed Abstractions. It fails to account for the relation that obtains when a list implements a stack.

On the other hand, since individuals and lower-level instantiables *can* be viewed as *implementations* of higher-level instantiables or universals, I suggest that instances and individuals are implementations, but not conversely.

7.5.2 Implementation as Instantiation.

Michael V. Anthony (1991) has explicitly argued that computer "implementations" are *not* "instantiations". The background of Anthony's argument is the "Classical" vs. "Connectionist" controversy, in particular, the question whether "a Connectionist architecture *instantiates* the Classical framework" or whether "a Connectionist architecture *implements* a Classical architecture" (p. 325, my italics), or whether there is some other (or no) relation between them.

As Anthony uses the term,

> 'Instantiation' expresses a simple relation between individuals and properties: an individual i instantiates a property P if and only if Pi. ... In the case of instantiation ... a *single* model or architecture is involved, and what is in question are its properties. (p. 325.)

Note that this is *not* necessarily the relation of instantiation we just looked at. There, instantiation was a relation between an individual or a category and its immediate superordinate category. Here, it is a relation between an individual and a *property*. Others have called the latter relation 'exemplification', though, of course, the relation between an individual and its properties can be (and has historically been) explicated in terms of category membership, and "exemplification" is the term often used for the relation between a real object and the Platonic Form that it "participates" in. Since I am not going to try to resolve several thousand years of metaphysics here, let's stick with Anthony's definition for now.

So, for a connectionist architecture to instantiate a classical framework would be for it to have classical properties. Let P_{cl} be the set of properties that "define the Classical ... framework" (p. 325). Let C_x be "a particular Connectionist architecture". Suppose that C_x has all of P_{cl}. Couldn't we then identify the Abstraction ClassicalSystem as the set of properties P_{cl} and treat C_x as an implementation of it? Anthony observes in a footnote (p. 339n7) that "individuals" could be "*abstract* objects like functional architectures" (my italics); thus, C_x is *also* an Abstraction (thus, clearly, Anthony is not speaking the language of §7.5.1).

In contrast,

> Where *implementation* is at issue ..., *two* functional architectures must be considered. A functional architecture FA1 is implemented, if at all, by the execution of a program in a distinct functional architecture FA2. (p. 325.)

So, FA2 might *itself* not have FA1's properties (so FA2 need not be an *instance* of FA1), but the *process*—the program in execution—*might* have FA1's properties (and so be an instance of FA1). In general, this seems OK. For instance, a machine-language program might have FA1's properties, but the machine language *itself* might not. As a trivial example, a machine-language program can have records, while the machine language itself doesn't have them.

"Intuitively," Anthony tells us, "the primitive operations, representational structures, etc. of FA1 get 'made up' or 'constructed' out of the resources of FA2. ... This is the relation that typically exists, for example, between assembly language functional architectures ... and higher-level architectures like LISP or Pascal ... when the latter are up and running on a computer"

(p. 325). So the idea is this: If FA2 (e.g., the machine language) has records as a primitive data type, then it's easy to implement FA1 (e.g, Pascal) in it, because they both already share the same properties—they both instantiate "having the record data type". If FA2 *lacks* records, they can nonetheless be implemented in it. But wouldn't FA2 then *have* records? Anthony seems to be trying to distinguish between essential properties and accidental ones: Records are an "essential" feature of a programming language that has them among its primitive (or built-in) data types, an "accidental" feature of a programming language that can only define them in terms of (or construct—or implement!—them out of) its primitive ones. (The latter are akin, too, to the notion of *derived* rules of inference in a natural deduction system.)

Consider, by way of analogy, the rationals and the integers. As is well known, the rationals can be ... implemented in? constructed out of? ... (equivalence classes of ordered pairs of) integers.[4] We'll come back to this implementation or construction or reduction in §7.5.3. For now, suppose we have a *language* for talking about the rationals (and sets). It will have, as one of its terms, a representation for $\frac{1}{2}$ (i.e., the rationals have, as one of its data objects, $\frac{1}{2}$). The language will also have as a primitive predicate some expression for a property P that applies only to rationals (i.e., the rationals have, as one of their properties, P)—e.g., (some) rationals have the property of being less than 1, and the set of rationals itself has the property of being dense. Now, suppose we have a language for talking about the integers (and sets). Can we talk about the *rationals* in the language of integers? Yes—by finding or constructing (analyzing?) the rationals' data objects (like $\frac{1}{2}$) in the integers and by finding or constructing (or analyzing) the rationals' properties among (or from) the integers' properties. By defining new *terms* in the language of integers, that language would now have terms for $\frac{1}{2}$ and P. That is, we could now say more in the language of integers than we thought we could; it *wasn't*, after all, limited to talking about integers. What, then, would be the difference between the language of rationals and the (extended) language of integers? The former would have certain terms and predicates (NPs and VPs) that the latter would lack; but they could be *defined* in the language of integers.

That this really is close to what Anthony has in mind can be seen from the following passage:

> ...in cases of *implementation*, lower-level architectures typically do not *instantiate* the *characteristic* properties of higher-level ones. An assembly-level architecture *implementing* LISP, for instance, does not also *instantiate* LISP: it lacks the *necessary* primitive properties (e.g., CAR, CDR), and has primitive operations LISP lacks (e.g., various operations on the contents of the accumulator). (p. 326, my italics.)

Here, 'characteristic' and 'necessary' can be taken to mean "essential". But doesn't a machine-language implementation of Lisp *have* `car`?[5] Maybe not: It can "simulate" `car`—or implement it?—but it doesn't *have* it; it can do what `car` does without *having* `car`. If you'll excuse the pun, I can do what can be done with a car without having one—by walking, taking the bus, etc.

We can draw a distinction between "weak" and "strong" implementations. For instance, a strong implementation of Lisp in machine language would be such that the machine language actually *had* identifiable data structures and procedures corresponding to lists, `car`, etc. A weak

[4] And vice versa, don't forget!—since $\mathbb{Z} \subset \mathbb{Q}$.

[5] `car` (or `first`) is the Lisp function that takes a list as input and returns its first member; `cdr` (or `rest`) is the Lisp function that takes a list as input and returns the "rest of" that list, i.e., the list consisting of all but that first member.

implementation of Lisp in machine language would be such that it would do the same things (e.g., be able to return the first element of a list) without having lists or `car` (just as I can get from my home to a store by car or by walking).

Conversely, "it is also true that an instantiation of LISP need not implement any distinct, higher-level LISP architecture" (p. 326). For example, I suppose, Franz LISP (understood as an *instantiation* (rather than an *implementation*?) of Lisp) need not *implement* SNePS. So, instances and implementations (as Anthony defines them) "are mutually independent" (p. 326).

Let's return to the rationals and the integers. Consider the integers first (and, for convenience, consider only the non-negative integers). What are they? One way to answer this is to cite Peano's axioms. That would be to present the Abstraction Integers—in fact, it is an abstract data type. Another way to answer the question is to say that integers are any things that satisfy Peano's axioms. So, e.g., the sequence consisting of $\emptyset$, $\{\emptyset\}, \{\{\emptyset\}\}$, etc., are integers. So is the sequence $\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}$, etc. So is the sequence $\emptyset, \{x \mid x \text{ is a set } \& \ \mathit{Cardinality}(x) = 1\}$, $\{x \mid x \text{ is a set } \& \ \mathit{Cardinality}(x) = 2\}$, etc. So is the sequence of symbol types 0, 1, 2, ..., 10, 11, 12, ..., 99, 100, etc. So is the sequence of symbol types 0, 1, 10, 11, 100, etc. (And if you ignore 0, so is I, II, III, ..., X, etc.). And so on. Note, for future reference, that the *differences* between each of these are what might be called "implementation details".

The vague feeling of discontent that this leaves us with is that there are *too many* integers, so to speak; isn't there an "intended interpretation"? What model-theoretic semantics teaches us is that there isn't. For any set of axioms, there are infinitely many models, including non-isomorphic ones. So, the only way to talk about "the" integers is to restrict ourselves to talk about Peano's axioms. (Recall the quotation from Russell, §3.2.1. Mathematics, on his view, is pure syntax.) The alternative is to choose, arbitrarily, some model of them and talk about *it*.

Now, since such axioms are abstract data types, such models are *implementations* in various media. For our first three examples above, the implementing media are sets "put together" in different ways; for the others, the implementing media are certain symbol types. If we restricted ourselves to some finite initial sub-sequence of the non-negative integers, we could take arbitrary physical objects as our implementing medium.

Are any of these implementations "instances" of ... integers? Peano's axioms? 'Instance' in the Porphyrian-tree sense? In Anthony's sense? I'm inclined to say 'No', but that's primarily because I find the interpretation of 'implementation' in terms of semantic models to be more illuminating. Moreover, I suspect that if one wanted to force the concept of an implementation into the mold of "instantiations", one could do so only by seeing "instantiations" as a kind of semantic modeling.

7.5.3 Implementation as Reduction.

Given all this, let's now consider the rationals (in particular, the non-negative rationals). Consider a set of axioms for the rationals; i.e., consider the abstract data type (NonNegative)Rationals. What are some of its implementations? Well, there are the fractions, i.e., the symbol types $\frac{0}{4}, \frac{1}{2}, \frac{2}{3}, \frac{1}{7}$, etc. There are the repeating decimals, i.e., the symbol types $0.\overline{0}$, $0.5\overline{0}$, $0.\overline{6}$, $0.\overline{142857}$, etc. There are also certain *constructions* from the integers, e.g., certain ordered pairs of integers: $\langle 0, 4\rangle$, $\langle 1, 2\rangle$, $\langle 2, 3\rangle$ $\langle 1, 7\rangle$, etc. Each of these can be considered to be an *implementation* of the rationals. Rational numbers are anything that satisfy the axioms.

Here, the notion of implementation details plays a larger role, since we seem to have "too many" rationals. In a Morning Star/Evening Star sense, $\frac{2}{3}$ and $\frac{4}{6}$ are the "same" rational number, as are the ordered pairs $\langle 2,\ 3\rangle$ and $\langle 4,\ 6\rangle$. We could say that that's an implementation detail, and provide rules (further axioms?) to indicate when two "intensionally distinct" rationals are "extensionally equivalent". We have such rules for, say, addition of integers: Does '$2+2=4$' state a fact about addition, or does it assert an extensional equivalence between intensionally distinct integers?. Or else we could—as in fact we normally do—implement the rationals as *equivalence classes* of ordered pairs of integers.

Now, often this "implementation" of rationals by integers (plus set theory), is called a "reduction" of the rationals to the integers. "All we really need," so the reductionist says, "are the integers (and set theory); we can define the rationals in terms of them (or, we can reduce the rationals to them)." So: Is implementation just reduction? Are all reductions implementations?

Again, we have related, but, I think, distinct, concepts. As Smith puts it, "*Reducibility* ... is a relation between *theories*; one theory is reducible to another if, very roughly, its predicates and claims can be translated into those of another" (Smith 1991: 280n39). Now, in the case of the rationals and the integers, I would really hesitate to say that the former have been "reduced" to the latter. I would be willing to say that the *theory* of rationals can be reduced to the *theory* of integers-plus-sets. But even here, when we prove some theorem about rationals, we haven't proved a theorem about integers but, at best, about certain *sets* whose "ground elements" are integers. For example, to prove a theorem about the *rational* number $\frac{1}{2}$ would be to prove a theorem about the following arcane set of sets of integers and sets of integers: $\{\{a,\{a,b\}\} \mid a,b \in \mathbb{Z}^{+} \ \&\ 2a=b\}$.[6] Suppose integers are implemented as sets, and multiplication is implemented as a set of ordered pairs of factors. Then we might have the following situation: If $\{\{\}\}$ and $\{\{\},\{\{\}\}\}$ implement 1 and 2, respectively, then $\frac{1}{2}$ could be implemented as the monstrosity $\{\{a,\{a,b\}\} \mid a,b \in \mathbb{Z}^{+} \ \&\ \{\{\{\},\{\{\}\}\},\{\{\{\},\{\{\}\}\},a\}\}=b\}$. The mind boggles. This is supposed to be *easier* to understand than '$\frac{1}{2}$' or '$0.5\overline{0}$'?

And the only reason we're interested in those rather arcane sets "of" integers is because they implement—are models of—the abstract data type Rationals. We might feel more "comfortable" with these arcane sets insofar as we are more comfortable with good old-fashioned sets and integers rather than with rationals *per se*. But that is an epistemological consideration that is rather suspect in the long run.

Once we have implemented the rationals using integers and sets, we also have another implementation of the integers, of course (since the integers are a proper subset of the rationals—or perhaps it would be better to say that a certain proper subet of the rationals is an implementation of the Integer abstract data type). I have in mind here the sequence $\frac{0}{1}, \frac{1}{1}, \frac{2}{1}, \frac{3}{1}$, etc. As a matter of fact, there are several implementations of the integers to be found among the rationals; here are a few:

$$\frac{0}{2}, \frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \ldots$$
$$\frac{0}{3}, \frac{1}{3}, \frac{2}{3}, \frac{3}{3}, \ldots$$
$$\frac{0}{1}, \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots$$

[6] The ordered pair $\langle 1,\ 2\rangle$ "is" (or can be implemented as!) $\{1,\{1,2\}\}$. The equivalence class containing $\langle 1,\ 2\rangle$ "is" $\{\langle a,b\rangle \mid a,b \in \mathbb{Z}^{+} \ \&\ 2a=b\}$. So, the rational $\frac{1}{2}$ "is" $\{\{a,\{a,b\}\} \mid a,b \in \mathbb{Z}^{+} \ \&\ 2a=b\}$.

These (especially the last one) may seem a bit odd, but recall that Peano's axioms only require that there be a successor relation, not that that relation be (implemented as) +1 (or even as +Successor(0)); the choice of a 0-element is arbitrary, too. Other arbitrarily strange ones are possible, e.g.,

$$\frac{0}{1}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots$$

And, of course, using an order imposed by a diagonalization, the (non-negative) rationals themselves can be taken as an implementation of the integers; e.g., if we arrange them (ignoring equivalences) two-dimensionally as follows:

$$\frac{0}{1}, \frac{0}{2}, \frac{0}{3}, \frac{0}{4}, \ldots$$

$$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$$

$$\frac{2}{1}, \frac{2}{2}, \frac{2}{3}, \frac{2}{4}, \ldots$$

then the sequence $\frac{0}{1}, \frac{0}{2}, \frac{1}{1}, \frac{0}{3}, \frac{1}{2}, \frac{2}{1}, \frac{0}{4}, \frac{1}{3}, \frac{2}{2}, \ldots$ implements the integers, too.

Finally, we could, if we wanted to, *re-implement* the rationals in one of these implementations of the integers, by the usual ordered-pair construction.

Why bother? Well, besides whatever insights such playful model-making gives us into the logical structure of the integers, it also shows that *reduction* (or construction) for the purposes of providing stronger epistemological foundations is *not* what implementation is. All of the above are implementations; none serves any interesting or useful reductive purposes.

The upshot is that although some, or even all, reductions or constructions might be implementations, certainly not all implementations are reductions.

7.5.4 Implementation as Supervenience.

There is one major concept that has a fairly precise definition and that is a good candidate for interpreting implementation, namely, supervenience. Recall, first, where we stand with respect to implementation, semantics, and syntax: An implementation of an Abstraction in some medium is a semantic model of the Abstraction in the "medium" of some semantic domain. And a semantic model is *any* structure—including the Abstraction itself!—that can be correlated (or put into correspondence) with the Abstraction. The closer the correlation, the better the semantic interpretation, even if, in the base case of a *self*-interpretation, we must resort to syntactic understanding.

Supervenience is certainly a plausible candidate. As Smith notes, "the term *supervenience* is used to relate phenomena themselves; thus the strength of a beam would be said to supervene on the chemical bonds in the constitutive wood. ... [S]upervenience doesn't necessarily imply reducibility" (Smith 1991: 280n39). When one domain "supervenes" on another, is it *implemented* by that other domain? And when one domain is *implemented* by another, does it supervene on that other domain? What, then, is supervenience? To answer this, let's look at several of Jaegwon Kim's classic papers.

7.5.4.1 Supervenience: An introduction.

In "Supervenience and Nomological Incommensurables" (1978), Kim gives a precise formulation of the informal notion that "one *family of properties* is 'supervenient' upon another *family of properties* in the sense that two things alike with respect to the second must be alike with respect to the first" (p. 149, my italics). The ultimate goal is to see whether it makes sense to say that mental properties are supervenient on physical ones. Consider, for the moment (we'll refine this later on), the rough analogy of mental properties to software and physical properties to hardware processes. If we have two like processes, do we have two like mental properties? This depends on how much "alike" the two processes must be: (a) They might be identical (same hardware, same machine-language level, etc.), or (b) they might be two different implementations with the same algorithmic behavior.

Consider (a): Could two identical processes be the result of distinct programs? Yes, *if* it's possible to have two distinct high-level–language programs that compile into the same machine-language program. Any other possibility (e.g., two distinct machine-language programs) would have processing differences. For instance, consider two (high-level–language) programs that differ only in that one of them has a bunch of "no-ops":

Program 1:

```
begin
  x := 2;
  y := 3;
  z := x + y
end.
```

Program 2:

```
begin
  x := 2;
  for i := 1 .. 100 do begin end;
  y:= 3;
  z := x + y
end.
```

Program 2 will run more slowly. With a suitable optimizing compiler, however, both could compile into the *same* machine-language program. So, in this case, the "mental" would *not* supervene on the "physical", *except* in the sense that the two high-level programs have the same input–output behavior and the same algorithm (except for eliminable or non-essential differences).

Consider (b), different implementations with the same algorithmic behavior: For instance, consider, first, Cassie implemented on a Sun workstation running SNePS in Allegro Common Lisp using the representations of Shapiro & Rapaport 1987 vs. Cassie implemented on a TI Explorer running SNePS in TI Lisp using the (different) representations of Wyatt 1989, 1990, 1993. Assume, further, that their input–output behavior is identical. But *do* they have the same *algorithmic* behavior? Arguably not, because of the representations (this, at any rate, is an empirical question;

cf. §5.5). So consider instead two Cassies as before, but both using the Shapiro & Rapaport 1987 representations. In this case, any differences would be *implementation* side-effects, *not* "mental" differences.

The upshot of all this? So far, given several plausible assumptions and the informal presentation of supervenience, the mental does seem to be supervenient on—as well as implemented by—the physical.

The other informal aspect of supervenience is that "there is no relationship of definability or entailment between the two families" of properties (pp. 149–150). Now, implementation is neither definability nor entailment, so it could indeed be supervenience. And, as Donald Davidson points out, "Dependence or supervenience of this kind [viz., of the mental on the physical] does not entail reducibility through law or definition" (Davidson 1970: 88; cited in Kim 1978: 150). As we saw in §7.5.3, not all implementations are reductions, so implementation—like supervenience—does not entail (hence, is not) reducibility; a program isn't "reduced" to a process, nor is a program defined in terms of the process (except maybe in cases of "reverse engineering", where we seek to discover in a bottom-up fashion *what* a computer is doing by examining the details of its behavior).

Now, according to Kim, "the main point of the talk of supervenience is to have a relationship of dependence or determination between two families of properties *without* property-to-property connections [or "correlations"] between the families" (p. 150). But in the case of implementation there *are* such property-to-property correlations. So maybe implementation *isn't* supervenience? As we will see, however, Kim's explication of supervenience allows for such correlations.

Note, by the way, that it's not a question of whether mental properties "emerge" from physical properties in some mystical way. Given a program and an implementing process, *there are* correlations between them: Suppose we have a process whose behavior can be described (from the intentional stance?) in mental language. Here, we would be giving a *functional* characterization of the process—what it does and how it does it, "modulo" its physical description (cf. Goguen et al.'s quotient spaces). Suppose, then, that we are able to do this sort of reverse engineering for some physical process. We would then have the mental and the physical, and could have the correlations.

However, Kim distinguishes between supervenience and functionalism/implementationalism (p. 150, col. 2). His point is that he's more interested in the fact "that there seem to be mental states which are 'nomologically incommensurable' with respect to neurophysiological or, more generally, physical properties; there appear to be mental states which do not nomologically correlate with physical states" (p. 150). The *reason* for this nomological incommensurability is not of interest to him. So, let's move on to other issues.

The fundamental intuition is that "what happens at the psychological level is fixed in every detail once the neurophysiological events are fixed. If you recreate Mr. Jones molecule by molecule, atom by atom, then the replica will have the same mental life that Mr. Jones has" (p. 151).[7] Note also that this is a strictly internalist, methodologically solipsistic approach—there's no talk here of external relations.

Kim first defines two set-operations, $^{\#}$ and * (p. 152, col. 1):

cite defs of $\mathbf{M}^{\#}$ and $\mathbf{M}^{*}$

[7]Kim compares this to *Star Trek*. On this, see Blish 1970, esp. Ch. 1. And on molecule-by-molecule "re-creation", cf. Castañeda 1989b and § **[on C*], [???]** below.

Consider an example. Let P, Q **[roman font ???]** be properties, and let $M = \{\mathrm{P}, \mathrm{Q}\}$. Then $M^{\#} = \{\mathrm{P}, \mathrm{Q}, \mathrm{P} \wedge \mathrm{Q}, \mathrm{P} \vee \mathrm{Q}, \mathrm{P} \rightarrow \mathrm{Q}, \neg\mathrm{P}, \neg\mathrm{Q}, \neg(\mathrm{P} \wedge \mathrm{Q}), \neg(\mathrm{P} \vee \mathrm{Q}), \ldots\}$, and $M^{*} = \{\mathrm{P} \wedge \mathrm{Q} \wedge (\mathrm{P} \vee \mathrm{Q}) \wedge \ldots,$ $(\mathrm{P} \wedge\neg\mathrm{Q}) \wedge (\mathrm{P} \vee\neg\mathrm{Q}) \wedge \ldots, \ldots\}$, where each element of M^{*} is an M-maximal property (and—in our example—the first-listed element of M^{*} contains no occurrences of $\neg$P or $\neg$Q and the second-listed contains no occurrences of $\neg$P).

Next, let D be a domain of objects, and let M, N be sets of properties that elements of D can have. Then *M is supervenient on N with respect to D* $=_{df}$ $\Box$(objects in D that share all properties in $N^{\#}$ also share all properties in $M^{\#}$) (p. 152, col. 1). That is, Suppose M supervenes on N with respect to D, and let $d, d' \in D$. Then $\Box(d, d'$ share all properties in $N^{\#} \rightarrow d, d'$ share all properties in $M^{\#})$.[8] What's meant is not that d, d' *have* all properties in $N^{\#}$, but that *if* they have all and only the *same* properties in $N^{\#}$, then they also have the same properties in $M^{\#}$. So, where D, d, d', N, M are as before, *M supervenes on N with respect to D* $=_{df}$ $\Box((\forall\ \mathrm{P}_N \in N^{\#})[\mathrm{P}_N(d) \leftrightarrow \mathrm{P}_N(d')] \rightarrow (\forall\ \mathrm{P}_M \in M^{\#})[\mathrm{P}_M(d) \leftrightarrow \mathrm{P}_M(d')])$.

Now suppose that M is a set of mental properties and N is a set of physical properties. What might D be? What kind of thing has both mental and physical properties? Descartes would probably say that *nothing* has *both* sorts of properties—that a *res cogitans* is not a *res extensa*. But perhaps a "person" or a "self" or even an AI computer (process) has both? Consider a computer programmed to compute greatest common divisors. We can equally well say, giving a sort of "mental" description, that it computes greatest common divisors and, giving a "physical" description, that it has a register that stores certain numbers (etc.). So, what the supervenience of the mental on the physical says is that if, say, two computers or two persons have the same physical properties, then they have the same mental properties, *where there is not necessarily any relevant or interesting relation between M and N*—it's just a correlation, but not (necessarily) a semantic one, since it's not (necessarily) point-by-point—there need be no patterns to match.

What, then, of implementation? Kim presents an argument that reducibility and definability *do* (*pace* Davidson?) entail supervenience (p. 152, col. 1). Can we run a similar argument to show that if M is *implemented* by N, then M supervenes on N? The argument requires biconditionals between N and M. Surely, if N implements M, there are such biconditionals. They would be provided for by the semantic interpretation function between N and M. Suppose that two things diverge on some M-property. Then they'll diverge in $N^{\#}$. So, if there are such biconditionals, then implementation does entail supervenience. Are there really such biconditionals? Since N implements M, there could be implementation side-effects (the domain of semantic interpretation might be "bigger" than the image of M in it). Still, if things diverge on M, they'll diverge in $N^{\#}$ (though perhaps not conversely; cf. the discussion of qualia in §7.6.3).

Kim argues, *pace* Davidson, that supervenience on a *finite* N entails that "each property in M which is instantiated is biconditional-correlated with some property in $N^{\#}$" and that such generalizations are lawlike (p. 152, col. 2). This is surely true for implementation in the N-to-M direction (p. 152, col. 1). Is it true in the M-to-N direction (p. 152, col. 2)? Suppose that $Q_1, \ldots, Q_n$ are the physical properties of the implementation, that P is an M-property, and that $Q_1 \vee \ldots \vee Q_n \rightarrow P$. Suppose, by way of contradiction, that x (e.g., a computer process) has P (e.g., a certain input–output behavior) but that x lacks each Q_i (i.e., is implemented differently). However, x is implemented *somehow*; let K be a property that x has in virtue of its implementation. Suppose y (some other process) also has K. Now, since M supervenes on N (N implements M),

[8] That, at least, is what Kim *says*; but doesn't he mean M^{*} and N^{*}? Perhaps not; cf. p. 153, col. 1.

y has P (i.e., y has x's input–output behavior). So, K must be one of the Q_is. Kim's argument seems to carry over (although details of the relationships between M, N and P, Q are not clear).

Moreover, supervenience *is* a semantic relation:

> To summarize: (1) if M supervenes on N, there are property-to-property correlations between M and N; (2) every property in M has either a necessary or sufficient condition in N ...; (3) if N is finite, every property in M is biconditional-connected with some property in N. ... [F]inite-based supervenience ... guarantees for each property in the supervenient family a co-extension in the supervenience base; and depending on the modality that attaches to the correlations between the two sets of properties, this may yield reducibility and definability. (Kim 1978: 153–154.)

Viewing the supervenient set as the mental realm and the supervenience base as the physical realm, each mental property has a co-extensive physical property and might be reducible to it, or definable in terms of it. The co-extensiveness *almost* works for implementation, but, strictly speaking, it doesn't. For the implementing device is not a set of properties; hence, it has no extension. Rather, it *is* the extension of the mental (or Abstract) properties. It is an open question what the appropriate "modality" is for sets of mental properties and sets of physical properties.

Kim asks, "What is the basis of our belief that, say, metaphysical processes wholly determine all other processes?" (p. 154). We might ask, more generally, about the basis of our belief that supervenience-base properties determine supervenient properties and, more specifically, that physical properties (of the implementation) determine mental properties (Abstract properties). But this latter question doesn't sound right for implementations. It sounds right for some kinds of epiphenomenalism, or for an identity theory, or for a theory that mind is an emergent property of the brain, etc., but not for implementations. One reason is that we don't think of an implementation as *determining* an Abstraction. Rather, the Abstraction is epistemologically prior; this is one of the early functionalist methodological principles (cf. Chomsky 1968: 12, Fodor 1968). One can, of course, "infer" an Abstraction ("abstract" it) from an implementation; but that is (merely) reverse engineering.

Finally, however, there is a problem in assimilating implementation to supervenience: Implementation isn't a relation between sets of properties. It's a relation between "physical" things and Abstractions—a relation between two different *kinds* of things—whereas supervenience is a relation between sets of properties.

7.5.4.2 Kinds of causation.

What about causation? I have argued elsewhere that, *pace* Searle, there are a variety of causal (or quasi-causal) relationships wherever there is an Abstraction, an implementing medium, and an implementation of the former in the latter (Rapaport 1985b, 1988a). To summarize briefly: Searle (1984) distinguishes only between what I call the Abstraction and the implementing medium, saying that the latter "causes" and "realizes" the former and that there are causal relationships within the Abstraction and within the implementing medium. As I see it, however, between the Abstraction and the implementing medium is the implementation itself (just as the computer process is between the program and the computer). Recalling the analogy of plays, Hamlet is an Abstraction, Olivier (say) is the implementing medium, and Olivier-playing-Hamlet is the implementation of the former by the latter. This yields a family of causal, quasi-causal, and implementation relationships:

1. The relation between the Abstraction and the implementation. (In Rapaport 1988a, I said that an Abstraction was a "species" of an Implementation, and an Implementation was an "instance" of an Abstraction. In view of my remarks in §7.5.2, above, this perhaps should be taken with a grain of salt.)

2. The relation between the Abstraction and the implementing medium (called 'realization' in Rapaport 1988a, following Searle's terminology, though not his relata). Again, in view of my remarks thus far in this chapter, I am no longer sure that this terminology is accurate.

3. A relation of ordinary, physical causation within the implementing medium (assuming, for now, that the implementing medium is a physical one).

These are the basic relationships. Definable in terms of them, we have:

4. The relation between an implementation and the implementing medium. (The medium "realizes" a "species" of the implementation; conversely, an implementation is an "instance" of that which is "realized" in the medium.)

5. A quasi-causal relationship among events and objects *within* the implementation: For example, Olivier-playing-Hamlet "causes" Jean Simmons-playing-Ophelia to do certain things: Olivier-playing-Hamlet is an "instance" of Hamlet, which is "realized" in Olivier, who, in turn really causes real events in Simmons, who "realizes" a "species" of Ophelia.

6. A quasi-causal relationship among events and objects *within* the Abstraction: For example, Hamlet causes Ophelia to do certain things because Hamlet is "realized" in Olivier, who really causes real events in Simmons, who "realizes" Ophelia.

Whatever we choose to call them, however, these relationships are all there.

In "Causality, Identity, and Supervenience in the Mind–Body Problem" (1979), Kim considers the same sort of issue from the point of view of supervenience. Consider psychophysical causation: how mental events cause brain events, and vice versa. Compare this to what might be called "algorithmic–physical" causation: how programs cause computers to behave.[9] Well, how do they? When the program is compiled, certain "switches" are set—i.e., the contents of certain registers are set to certain strings of 0s and 1s. This is done according to machine-language specifications, which are translations from the higher-level programming-language specifications. When the program is executed—when the *process* comes into being—the switch-settings permit the flow of energy to reset some of the switches, continually, until the program halts (i.e., until the computer's switches reach a state in which none of them permit any more energy-flow to reset any of them). Similarly, mind→brain "causation" ought to consist of physical changes to brain states—the analog of switch settings. How would brain→mind "causation" work? Well, how does computer→program causation work? By switch settings that are interpreted as program, not data. An output program in such a situation would just be a symbolic record of the switch settings in

[9]And vice versa? It can happen when programs modify themselves.

the program-region of memory. So, in mind–brain interaction, perhaps the mind just *is* "switch settings" in the brain.

Kim sets the scene as follows:

> Let M be a mental event (type), and let P be its neural correlate. ... Assume further that M is a cause of a physical event P^* (this is the posited psychophysical causal relation), and ... that there is a law linking M with P^*. It follows that ... a law exists that links P and P^* .... Now we can see three related puzzles arise. (Kim 1979: 35.)

Let M be a piece of program code, let P be its corresponding switch settings, and assume that M causes some other switch settings P^* in a lawlike way. Then P and P^* are lawfully linked. But is it really M that causes P^*? It seems better to say that it is the *implementation* of M that causes P^*. P^* can be described in programming-language terms; call this description 'M^*'. What is the relation between M and M^*? I would say that it is the quasi-causal relationship 6, above.

The puzzle that is of most interest for our purposes among the three that Kim sees in this situation is what he calls

> *The problem of pre-emption.* Given that M and P are nomic equivalents and given that there is a law linking P with P^*, as well as one linking M with P^*, P appears to have at least as strong a claim as M to be the cause of P^*. (Kim 1979: 35.)

That is, P "pre-empts" M as cause of P^*. Kim suggests that psychophysical "causation" (the relation between M and P^*) is not "real" causation (*pace* Searle!) but a "simultaneous equivalence" (p. 35; cf. his earlier notion of "Cambridge dependence" (Kim 1974)). I would prefer to call it ... implementation. Note that even if it is not the case that M *causes* P, still M is crucial for understanding what P is and what its *function* is (what it does). So we needn't despair that psychophysical "causation" isn't what we might have thought it was, that—as Kim puts it, citing Norman Malcolm—"common-sense psychological explanation of bodily motions in terms of beliefs, desires, intentions", etc., would be pre-empted (p. 36). What I am taking to be implementation is what underwrites our ability to take the Dennettian intentional stance, what underwrites Chomskian/Fodorian functionalist methodology, and what makes reverse engineering difficult (knowing the switch settings without knowing the program doesn't easily tell us what a computer is doing).

Kim generalizes the problem to properties and events, so we could talk—instead of "programs" and "computers"—of *properties* of programs and *physical properties* of computers, or of *computational processes* (which are events) and *physical processes*. Do properties of programs (e.g., the property that a certain variable is declared as being of a certain type) cause physical properties of computers (e.g., that certain switches have certain settings)? Well, the program properties certainly *explain* the physical properties. Does it, for that matter, make sense to say that one *property* causes another? Surely, were the program properties different, the physical properties would be different (different programs produce different physical behaviors.) On the other hand, the computational *process* seems to be only a *different description* of the physical process, along the lines of the intentional stance.

So, where does supervenience come in?

> … if two organisms are metaphysically indistinguishable from each other, then they will share the same psychological life; and if two physical objects are metaphysically indistinguishable from each other, then they will share the same macro-properties. (Kim 1979: 40.)

Now, Kim also says (p. 41) that the macro-properties are *supervenient* on the micro-properties *even if the macro ones could have been "realized"* [p. 40] *in different ways*. This suggests that implementation is *not* supervenience. But the quotation suggests that two implementations will be implementations of the *same* Abstraction, which *seems* inconsistent with, for instance, the cases of a single computer program that is playing chess and re-enacting a Civil War battle, or a natural-language text that is both about chemical and mathematical lattices.

One way out of this might be to distinguish between the Abstraction and an interpretation, saying that in such cases there is only one Abstraction but several interpretations. However, it is the interpretation that I have hitherto considered to be the Abstraction.

Another way is to use Smith's distinction between the actual world, a mathematical model of it, and a program: The program implements the mathematical model (the Abstraction), but the mathematical model is only partial and *could*, therefore, be a model of lots of different aspects or parts of the actual world. Also, surely, lots of different parts of the actual world *have the same structure*—that is, the same mathematical model.

Now, Kim offers a slightly different definition of supervenience:

> A family M of properties is *supervenient upon* a family N of properties with respect to a domain D *just in case* necessarily, for every property P in M and each object x in D such that x has P, there is a property Q in N such that x has Q and any object y in D which has Q also has P. (Kim 1979: 42.)

So, taking D as the domain of, say, intelligent entities (cognitive agents), which can have both mental (M) and physical (N) properties, for each mental property $P_M \in M$ that cognitive agent $x \in D$ has, there will be a physical property $Q_N \in N$ that x has, and any other cognitive agent $y \in D$ that has Q_N also has P_M. This last clause, about y, suggests a third way out: The chess-playing computer *is* a Civil War computer; only the *interpretation* of its input and output differs. So the interpretation is not a mental property.

However, the chess properties are *also* supervenient on N. So what happens when two distinct Ms supervene on N? What would be the relationship between the two Ms? (See Figure 7.3.) Could it be that M_1 supervenes on M_2? Choose $x \in D$. Choose some property in M_1, $P_{M_{1_i}}$, such that $P_{M_{1_i}}(x)$. Then, since M_1 supervenes on N, there is $Q_{i_1} \in N$ such that $Q_{i_1}(x)$ and which is such that, for any $y \in D$ such that $Q_{i_1}(y)$, $P_{M_{1_i}}(y)$. Similarly, since M_2 supervenes on N, there is $Q_{i_2} \in N$ such that $Q_{i_2}(x)$, and that is such that, for any $y \in D$ such that $Q_{i_2}(y)$, $P_{M_{2_i}}(y)$. Now, in the case at hand, we can assume that $Q_{i_1} = Q_{i_2}$. The idea is that, say, the object Soldier-1 in the Civil War program is "implemented" by the same data structure that implements the object Pawn-1 in the chess program. To see whether M_1 supervenes on M_2, we need to see whether there is $P \in M_2$ such that $P(x)$ and which is such that, for all $y \in D$ such that $P(y)$, $P_{M_{1_i}}(y)$. The obvious candidate for P is $P_{M_{2_i}}$, which, by hypothesis, is such that $P_{M_{2_i}}(x)$. Under what circumstances would an arbitrary y that is such that $P_{M_{2_i}}(y)$ also be such that $P_{M_{1_i}}(y)$? Well, we know that there is $Q \in N$ that is such that $Q(x)$, namely, Q_{i_1} (i.e., Q_{i_2}).

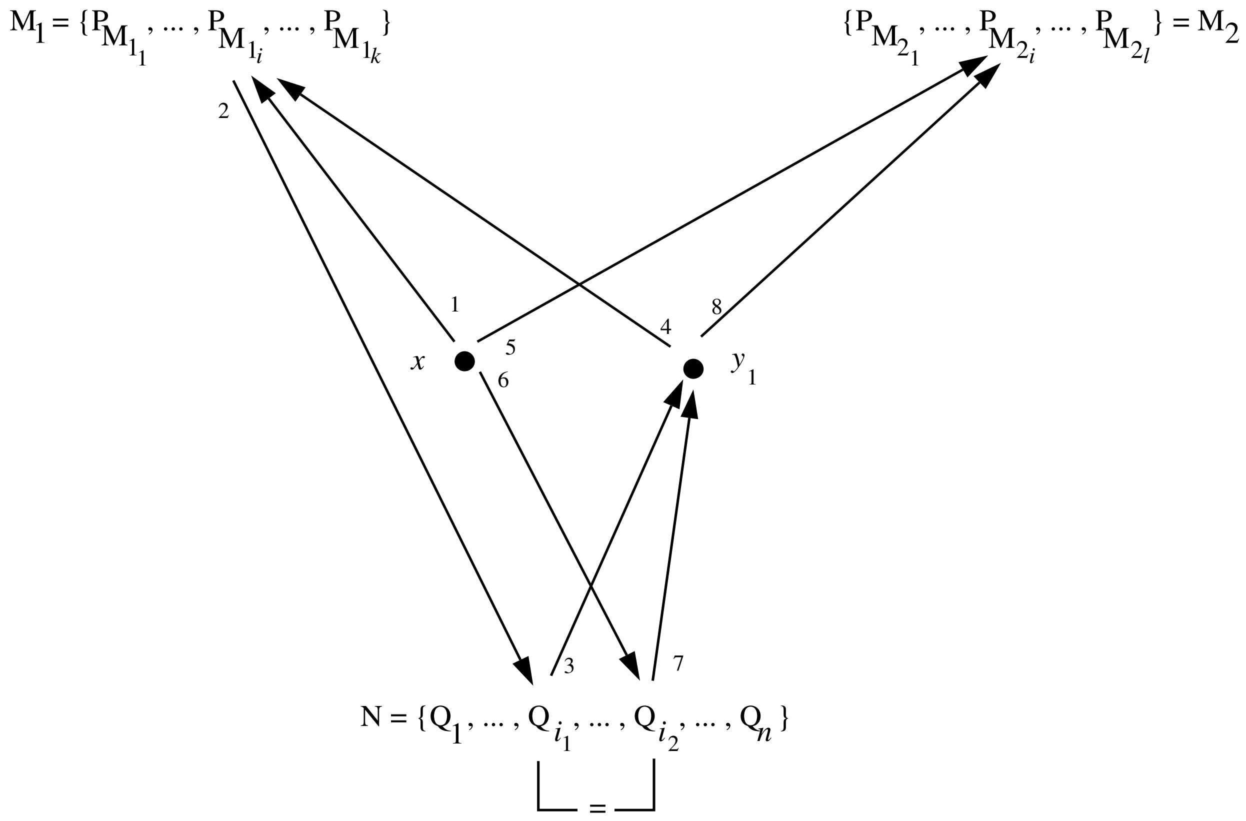


Figure 7.3: M_1, M_2 supervene on N. What is the relationship between M_1 and M_2?

And we know that an arbitrary y that is such that $Q_{i_1}(y)$ is also such that $P_{M_{2_i}}(y)$ *and* $P_{M_{1_i}}(y)$. So it appears that M_1 supervenes on M_2 (unless, perhaps, the 'necessity' clause rules that out). So, the chess properties and the Civil War properties appear to be mutually supervenient. But what does *that* mean? And there are further questions: Is supervenience symmetric? Transitive? (On these questions, see §7.5.4.3, below.)

And how does all of this relate to implementation? Let A be an Abstraction and I be its implementation in some medium S. Then it is S that has both A-properties and I-properties. So A-properties *could* supervene on I-properties. So, possibly, A is implemented by I if and only if A-properties supervene on I-properties. Indeed, Kim (pp. 43–44) sees supervenience as a very general version of the family of concepts that includes reducibility, etc. So perhaps it is the base relation in terms of which the others can be defined? I am uncomfortable with this, primarily because I see the generalized *semantic* relation as the fundamental one, and I take implementation to be a specific case of a semantic relation. So, too, for supervenience, which is, as we saw, a correlation relation.

Here is how Kim puts it:

> The macro-property of water-solubility may be exemplified in diverse molecular structures .... Certain gases may be water-soluble in virtue of their molecular structure M, and certain solids may be water-soluble in virtue of having a different molecular structure M*. Thus M will be the supervenient base of these gases' solubility in water, and M* will be that of these solids' solubility in water .... (Kim 1979: 45.)

I would say that water-solubility is *implemented in* diverse molecular structures. In any case, this sounds exactly like my Abstraction/Implementation relation. So, an implementation (Medium) has the properties that form the supervenient base for the properties of its Abstraction. Indeed, Kim says as much:

> Suppose that an occurrence of pain causes a limb withdrawal. ... This situation will be represented in the following way: This particular pain has a supervenient base, say, the excitation of certain neural fibers, although, as the functionalists argue, there may be no biconditional law correlating pain and this sort of neural base. (Thus, the notion of supervenient base corresponds roughly to the notion of "physical realization" often used by the functionalists.) (Kim 1979: 45.)

But I think it's better to say that the supervenient base is the *properties* of the implementing medium. But aren't the supervenient properties *and* the supervenience-base properties *both* properties of the medium? Is one set (the base) more "fundamental" than the other? Perhaps the properties of the supervenience-base are more "proper"—they are *properties*, whereas the supervenient "properties" are *attributes* (attributed by whom?—by external observers?).

Kim further describes the situation thus: Not only is a pain supervenient on a neural base, but the limb withdrawal is itself rather "Abstract" and supervenient on an anatomical/physiological base.

> ... there is a causal path from the supervenience base of the ... pain to this supervenience base of ... [the] limb withdrawal. It is also part of our account that

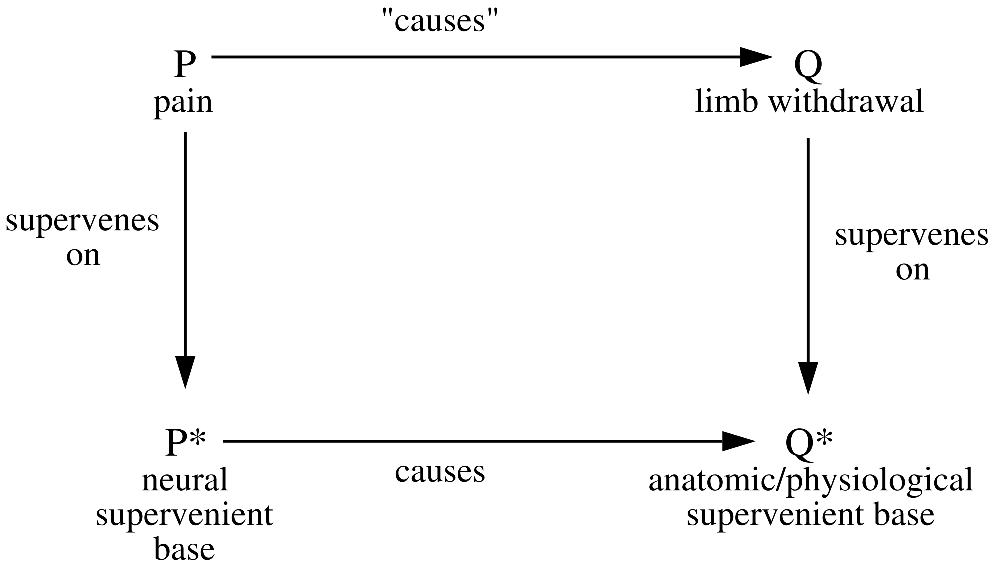


Figure 7.4: Kim's picture of the relationship among pain and limb withdrawal.

> the causal role of the pain vis-à-vis the limb withdrawal is *explained in terms of* this more basic underlying causal chain .... (Kim 1979: 45; my italics.)

Kim's picture is as shown in Figure 7.4. That is, the "causal" relationship between P and Q in Figure 7.4 is explainable (or definable) in terms of supervenience and physical causation. However, I recognize a *real* relationship at the top level, not merely a defined one; it may be *equivalent* to a combination of other relations, but it is *sui generis*. Where Kim says that "there is no independent causal path from ... P to ... Q or Q*" (p. 46), I would say that it *is* independent (though not really causal)—at the level of *functional* analysis. (Interestingly, in view of my remarks about good old-fashioned Cartesian dualism (§7.2), Kim calls this "supervenient dualism", and distinguishes it from epiphenomenalism.)

7.5.4.3 Supervenient causation.

Kim returns to causation in his 1983 essay, "Supervenience and Supervenient Causation". He begins by pointing out that one advantage of the notion of supervenience is that it allows us "to acknowledge the primacy of the physical over the psychological" (p. 45): The mental supervenes on the physical, but not vice versa (Kim 1982: 52), presumably because two entities could be in the same psychological state while in different physical states—i.e., psychological states can be multiply realized! Of course, if angels exist, then, presumably, they also think, so it's surely not the *physical* that we want to have "primacy" but—more generally—the implementing medium (cf. Fodor 1981).

But do we even want the implementing medium to have "primacy"? Do we want to say that nothing is or has a mind unless it is implemented? Would we want to say that nothing is an algebraic group unless it is implemented? In the case of minds, there does seem to be a desire for *behavioral processes*, hence for an implementation that can behave. But for the case of groups, there isn't. So, in general, both terms of the Implementation–Abstraction relation are of equal importance. One can't be *fully* understood (by a third party) without the other.

Kim defines 'weak supervenience' as before (p. 46; cf. §7.5.4.1, above). Note, again, that the object in the domain of the supervenience relation—the one that can have *both* the supervenient and the supervenience-base properties—*isn't* (merely) a computer or a physical object. It must be something (capable of) having *both* properties, such as a *person* (rather than a physical, human body) or a computer *process* (cf. Haugeland's objection, 1983: 65). A problem with supervenience is that there can be two "physically indistinguishable worlds" that are not also "psychologically indistinguishable" (p. 40). For instance, a computer c_w in world w running program p might be exhibiting some mental processes, but another computer $c_{@}$ in the actual world running p might not be. But is that so bad? Recall the chess/Civil War and chemical/mathematical lattice examples. In any case, Kim offers 'strong supervenience' as a remedy:

> A *strongly supervenes* on B just in case necessarily for each x and each property F in A, if x has F, then there exists a property G in B such that x has G, and *necessarily* if any y has G it has F. (Kim 1983: 49.)

and he points out that "Both relations are transitive, reflexive, but neither symmetric nor asymmetric" (p. 49). Transitivity is good; it's needed to account for levels of virtual machines, each of which can be said to supervene on, or be implemented by, a lower-level machine. Reflexivity, though, does not seem to be a property that we would want implementation to have. This means that *every* implementation *must* have implementation-dependent side-effects, since every implementation of an Abstraction will contribute something over and above what the Abstraction specifies.

If supervenience is non-symmetric, then it's possible for two properties to supervene on each other. Could each be an implementation of the other? That *seems* counterintuitive. surely, two things can implement each other—or be semantic interpretations of each other—but not at the same time. There is a directionality, a point of view of the third party that *uses* one domain as a semantic interpretation or implementation of the other—recall the discussions of the asymmetry of antecedent understanding (Chs. 2–4). So it looks as if supervenience is *not* implementation.

Is it reduction?

> If you believe that the mental strongly supervenes on the physical, you are committed to there being a physically necessary and sufficient condition for each psychological state. The physical base ... may not even be humanly discoverable; as a result it may be *unavailable* for a physicalist reduction or explanation of the psychological state. ...
>
> Thus, strong supervenience is not the same thing as the *reduction* of the supervenient family to the base family; reduction is an explanation procedure, and to carry out a reduction we must identify for each basic supervenient property its supervenient base property. (Kim 1983: 49–50.)

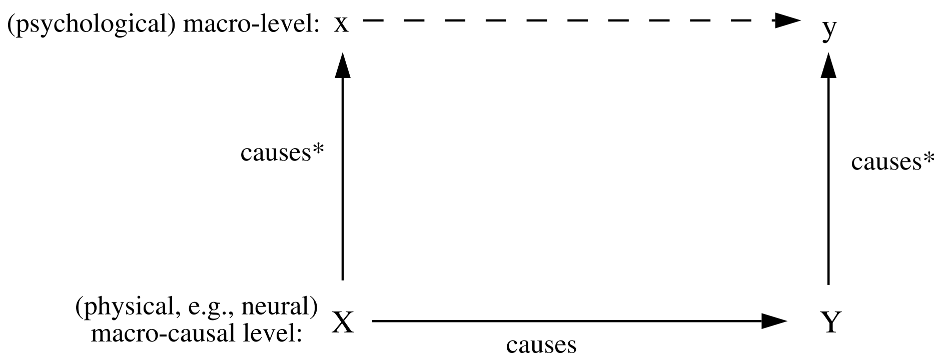


Figure 7.5: At the (physical) macro-causal level, physical state X really causes physical state Y. The relation between that level and the (psychological) macro-level—causation*—is really "mereological supervenience" (cf. p. 52).

But for an implementation, as for a reduction, we would need the identification, unless, perhaps, we have a certain kind of connectionist implementation. So, insofar as we *need* the identification for an implementation, implementation is *not* the same as supervenience. And insofar as we *don't* need the identification for an implementation, implementation is not the same as reduction. Independently, implementation is not reduction, because in reduction, one can jettison the reduced theory, but for implementation, one *needs* the Abstraction to tell you what's going on.

Kim introduces the notion of "supervenient causation" (pp. 50–51). Consider the situation shown in Figure 7.5. Here, the relation between states at the psychological level is what Kim calls 'supervenient causation', which he takes to be *reducible* to the *real* causation at the physical level. (Later (pp. 54–55), he also considers the relation between X and Y to be one of supervenient causation.) He points out that this is *not* eliminative materialism, because the psychological states are not "rejected". Nor is it Cartesian dualism, because supervenience is not causation. Nor is it epiphenomenalism, because supervenient causation is "as real as causal relations involving sundry macro-objects and their observable properties" (p. 55).

7.5.5 Summary.

So, what is implementation? We have seen that it is a very widespread phenomenon, taking many guises. It is a relation that obtains between two things—I have called them the Abstraction and the Implementation—when the Implementation is a "concrete" or "real" or "physical" thing that has all the properties of the Abstraction. But we have also seen that Implementations can be equally "abstract". So there are two sorts of implementations: abstract and concrete ones, the latter being "realizations" in some physical medium. We have seen that they typically have *more* properties than their Abstraction. So perhaps implementation is best construed (even etymologically) as a general term for *any* filling in of details; concrete implementations are fillings-in in concrete media. Thus, the notion of implementation comes along with a notion of "level": the more detailed level being "below" the "higher"—or more abstract—level, and the "concrete" or "physical" level being at the "bottom"—being the "foundation" as it were. Paul Smolensky characterizes it that way:

> For my own purposes, the crucial aspect of the implementation relation is this. Suppose we have a physical system S which at some ["lower- or 'micro'-]level of description L_μ is performing exactly the computation μ; that is, if we write down the laws governing the dynamics and interactions of those aspects of the system state that are characteristic of level L_μ , we find these processes to be exactly described by μ. If μ is an implementation of [a "higher- or 'macro'-level description] M, we are guaranteed the following: The states of this same system S have characteristics at a higher level L_M which evolve and interact exactly according to M: These characteristics define a description of S at the higher level L_M for which M is a complete, formal, and precise account of the system's computation.
>
> If μ implements M, then this constitutes the strongest possible sense in which μ and M could both be valid descriptions of the same system S. (Smolensky 1988: 59.) **[???]**

The one caveat I have here is that μ will typically have more details than M—it will say more about S, though, as we will see, what more it has to say may not be interesting from the point of view of M.

We have also seen that individuation, instantiation, reduction, and supervenience are all related to implementation, though "weaker" than it (cf., too, Smolensky 1988: 70n1). **[???]**

The single best "interpretation" of implementation seems to be that of semantic interpretation: I is an implementation of A in medium M if and only if I is a semantic interpretation or model of A, where A is some syntactic domain and M is the semantic domain.

In the remainder of this chapter, we shall look at some of the implications of this point of view: the role of the "implementation details", the question of whether an implementation is "the real thing", and the problem of whether anything can be an implementation of anything else.

7.6 IMPLEMENTATION-DEPENDENT DETAILS.

7.6.1 In the Details Lie the Differences.

Suppose we have two different implementations of an Abstraction. They may be implementations in different media, as, for example, implementations of the Stack abstract data type in Pascal using records and in Lisp using lists, or implementations of Cassie on a Vax running Franz Lisp and on a Sun running Allegro Common Lisp. Or they may be implementations in the *same* medium: for example, two implementations of a fully-equipped Ford Taurus LX—here what I have in mind is that the cars would be identical except, of course, that the metal, plastic, fabric, etc., of one of them would perforce be distinct physical objects from the metal, plastic, fabric, etc., of the other. For another example, consider two implementations of the Stack abstract data type using Pascal arrays; here, one implementation might use an n-dimensional array A with `top` $= A[n]$, while the other uses an $(n+1)$-dimensional array A with `top` $= A[0]$.

Clearly, the members of each pair of implementations will differ. The operations of the Pascal stack will be defined in terms, say, of record operations, while the operations of the Lisp stack will be defined in terms of list operations. The two Cassies' input–output behaviors ought to be the same, but the code will differ, so debugging will be a different process on each. One Taurus might dent more easily than the other, or get better gas mileage. (Even identical twins differ, as in Georges Duhamel's 1931 novel *Les Jumeaux de Vallangoujard*.) And clearly there will be implementation-dependent differences between the array-implemented stacks, some of which (e.g., "where" `top` is) are behaviorally (i.e., input–output) irrelevant and some of which (e.g., the size of the array) have behavioral consequences.

Jorge Gracia discusses the relation of an artist's "general idea of what he [sic] wants to do" and the final product, e.g., a sculpture:

> ... the sculptor's description is too general and does not identify those features of the sculpture that set it apart from others [that satisfy the description]. ... [t]he particular sculpture that the sculptor produces is not the result of his idea alone, but involves also the materials with which he works as well as the creative process itself that produces it. (Gracia 1990: 511–512.)

In general, then, when a given Abstraction is implemented in different media (or in different ways in the same medium), there will be implementation-dependent differences. Nevertheless, there will be some core, some essence, common to all of them in virtue of which they can be said to be the "same". This, I take it, is the point of Goguen et al.'s isomorphism construction.

Implementations are always more specific or detailed than their Abstractions, and they involve the implementing medium. This gives rise to *implementation-dependent side effects*. Similarly, ideas can be implemented in different languages (or differently implemented even in the same language). Clarity of exposition, literary art, and even cultural variety thrive on the implementation-dependent side effects due to the implementation-dependent differences. *Vive les différences!*

Consider the implementation of a mind. That is, suppose that (at some future time) we have a collection of algorithms that "account for" cognition—the Mind abstract data type, as it were. Suppose that we have neurological evidence that it is implemented in the human brain, and suppose that intelligence artificers (to use Dennett's happy term **SOURCE?**) have implemented it on a supercomputer. We should expect that there will be *implementation-dependent* differences between human minds and such computer minds. Does this mean that the computer mind is not a "mind"? I understand this question in the following way: Is the computer mind an implementation of the Mind abstract data type? The answer, by hypothesis, would clearly be 'Yes'. Are the differences "important"? That, of course, depends on what counts as being "important". Perhaps there will be a need to talk of *degrees* of "mindhood" (cf. Rapaport 1993b). Perhaps, for example, the Mind abstract data type will not be able to be fully implemented in dogs, or in chimps (or perhaps we will be able to distinguish between a Human Mind abstract data type and a Dog (or Chimp) Mind abstract data type).

Perhaps, in the long run, the *only* differences that will be of any significance will be the

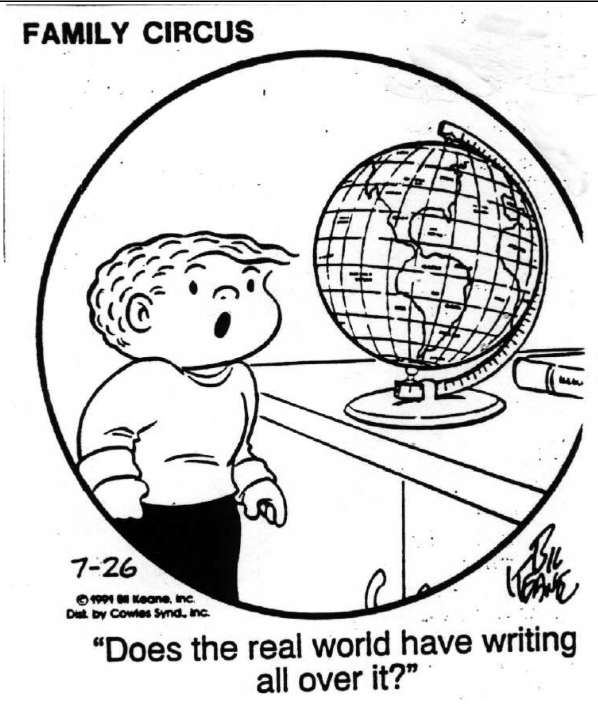


Figure 7.6:

implementation-dependent ones—the physical differences—and even these will be of no more (or perhaps no less) significance than the implementation-dependent differences that *currently* exist due to the fact that *your* mind is implemented in *your* body and mine in mine. Suppose, for example, that androids like Data of *Star Trek: The Next Generation* are commonplace. Would we—*should* we—behave differently towards them? Suppose, even, that some cognitive agents are "aware" or have "subjective experiences" (measured by, e.g., whether they have faces or are human, or by some primitive "feeling" or "intuition" that they are aware), while others are *not* thus "aware" (e.g., some computers). Suppose further that these two kinds of cognitive agents are not behaviorally distinguishable (perhaps only physically distinguishable—i.e., distinguishable on the basis of certain perceptual aspects of their implementation). Given this behavioral indistinguishability, I would say that we would *not* (better: should not; consider, after all, the ugly varieties of racism) behave differently towards them—not even towards the *non*-"aware" ones: For even they, because they were behaviorally indistinguishable, would *claim* to feel pain, say; so it would be morally wrong to inflict pain on them. What, then, would be the difference between them? Only a linguistic convention.

7.6.2 Implementation-Dependent Side Effects.

Consider an object that is a model of something. How much of it is part of its role as a model, and how much is due to its own nature—to its implementing medium? From the fact that a globe is plastic, we do not infer that the world is plastic. Nor do we infer that the world has writing on it from the presence of place names and lines of latitude and longitude on globes, even though these *are* part of its role as a model (Figure 7.6). Where do implementation-dependent side effects come from, and what, if anything, do they do?

Implementation-dependent side effects are due to implementation-dependent details. The details come from situations in which the semantic domain is "larger" than the syntactic domain. These are situations in which "completeness" holds in the sense that everything in the syntactic domain is interpreted in the semantic domain, yet in which not everything in the semantic domain is an interpretation of something in the syntactic domain. Thus, individuals can have properties that their universal lacks:

> ... written, spoken, and mental texts are all individual insofar as they are not instantiable themselves. ... As individual instances, moreover, they presuppose corresponding universals, but the universal is not the same for the three types of texts. *For the written text, it would be a written type of universal even though the universal would not be something written anywhere.* (Gracia 1990: 505–506; my italics.)

Such implementation-dependent properties, we see, *can* be essential properties of the individual; we'll come back to this in §7.6.3.

Another source of implementation-dependent details is non-isomorphic models. Recall the discussion in §2.2.2 of non-isomorphic models of the group axioms (i.e., of the Group abstract data type). For example, consider two groups of different cardinalities (e.g., the cyclic groups of orders 2 and 3) or an infinite cyclic group such as the integers under addition and the Cartesian product of that group with itself (which, unlike the former, has two disjoint subgroups except for the identity). In each case, the implementation—the model—has features that are left *unspecified* by the Abstraction (in this case, the group axioms); they are implementation-dependent details. Indeed, even *isomorphic* models give rise to implementation-dependent differences: "In any isomorphic class there are models which differ on *all* non-empty extensions. For example, in any isomorphism class there is one model at least whose domain consists of odd integers and one whose domain consists of even integers" (Jardine 1973: 231; Jardine points out that this gives rise to Quinean indeterminacy of reference).

Sometimes, the implementation-dependent details are not important and can—or even *must*—be ignored. This is because the *purpose* of an implementation or model is often to aid in understanding the Abstraction. There are two sides to this coin: If the *Abstraction*—better, the syntactic domain, the domain to be modeled, the domain to be understood in terms of the model—is itself complex, we will want the *model* to be simpler.[10] Nonetheless, it will still have features that do not represent any part of the Abstraction: "It may be *richer* in properties, but these would then not be ones relevant to its object [i.e., the Abstraction]; it [i.e., its object] wouldn't possess

[10] "It is rather paradoxical to realise that when a picture, a drawing, a diagram is called a model for a physical system, it is for the same reason that a formal set of postulates is called a model for a physical system. This reason can be indicated in one word: simplification. The mind needs in one act to have an overview of the essential characteristics of a domain; therefore the domain is represented either by a set of equations, or by a picture or by a diagram. The mind needs to see the system in opposition and distinction to all others; therefore the separation of the system from others is made more complete than it is in reality. The system is viewed from a certain scale; details that are too microscopical or too global are of no interest to us. Therefore they are left out. The system is known or controlled within certain limits of approximation. Therefore effects that do not reach this level of approximation are neglected. The system is studied with a certain purpose in mind; everything that does not affect this purpose is eliminated" (Apostel 1960: 15).

them, and so the model couldn't be taken to represent them in any way" (Wartofsky 1966: 6–7). The extra properties are implementation-dependent details, to be ignored. Again, recall the construction used by Goguen et al., to divide out such irrelevancies.

Often, however, the details *do* contribute something: This is the realm of the implementation-dependent *side effects*—phenomena contributed by the implementing medium, not by the Abstraction. Some are behaviorally relevant, others not. That a stack's top is implemented as $A[0]$ rather than $A[n]$ is not behaviorally relevant. A high-level program that cares only about stacks and not about their implementation can—and does—ignore this. In programming languages with built-in data-abstraction mechanisms, such as Modula-2, they *literally* ignore—do not know—about the implementation details (cf. Parnas 1972).

But as side effects become more and more behaviorally relevant, they become more than mere *side* effects and can be of central importance. For instance, in the chess game played with non-standard pieces (cf. §3.2.1), the implementation of the Abstract chess pieces had confusing implementation side effects. Sellars's Texas chess, played with Pontiacs implementing bishops (§7.4.2.2), will have, if not *confusing* side effects, certainly *significant* ones—the chess board will have to be pretty large, and perhaps a speed limit will have to be imposed on the bishops. More significantly, problems with an implemented computer system may be due to details of the implementation that are not part of the original specifications. That is, the system might mathematically "satisfy" the specifications, yet still fail due to hardware faults:

> ... hardware does from time to time fail, causing the machine to come to a halt, or yielding errant behaviour (as for example when a faulty chip in another American early warning system sputtered random digits into a signal of how many Soviet missiles had been sighted, again causing a false alert ...). (Smith 1985: 635.)

This, I take it, is at the heart of James H. Fetzer's arguments against program verification (Fetzer 1988, 1991; cf. Nelson 1992, 1994).

To some extent, the notion of an implementation-dependent detail and its attendant "side" effects is a relative one. Recall Gracia's example of the individual written text and its non-written "written type of universal". There would, however, be a further universal, of which the "written-type" and "spoken-type" of universals are instances. (For example, a high-level universal might be Lincoln's Gettysburg Address, of which the written-universal and the spoken-universal are species; one written individual falling under the former would be the one Lincoln allegedly wrote on the back of an envelope, and one spoken individual falling under the latter would be the one Lincoln uttered on 19 November 1863). Or compare Euclid's algorithm for computing greatest common divisors with that algorithm implemented in Pascal, and with that algorithm implemented in Lisp; each of these can be (further) physically implemented as processes on a variety of machines. Each level of Abstraction or implementation ignores or introduces certain details. One level's implementation detail is another's Abstraction. That is, we can (via a kind of reverse engineering) "abstractify" an implementation's details, after which they are no longer "details" *relative to* the new (more detailed) Abstraction. Consider, for example, the Stack abstract data type and the N-Element Stack abstract data type. A Pascal n-element array-implementation of a stack (*simpliciter*) will have as an implementation detail (yielding behaviorally observable side effects) that it can only store n elements. Yet the very same code will *also* be an implementation of an N-Element Stack and, as such, will *neither* have that feature as an *implementation*-dependent detail *nor* as a *side*-effect—indeed, it will be an essential feature.

Note that we have two senses of 'abstract' here: the sense in which abstract data types, specifications, and blueprints are "abstract" (relative to implementations) and the sense in which to abstract is to eliminate (or ignore) "inessential" "details": "every model deals with its subject matter *at some particular level of abstraction*, paying attention to certain details, throwing away others, grouping together similar aspects into common categories, and so forth" (Smith 1985: 637). Note, too, that the model need only be "*assumed* simpler" (Rosenblueth & Wiener 1945: 317; my italics): The implementation-dependent details are *ignored*, not *eliminated*. They are parts of the model that are *not* (intended to be) representations of the system being modeled.

7.6.3 Qualia: That Certain Feeling.

The view of implementation as semantic interpretation, with its implementation-dependent details giving rise to implementation side-effects, suggests a solution to the puzzles of qualia. Qualia, roughly, are the subjective, qualitative "feelings" or "sensations" or "experiences" that accompany various mental states and processes. Examples are the "look" of blue (as opposed to yellow, and of yellow as opposed to blue) and the "feel" of pain (or, for that matter, tactile sensation *simpliciter*). The puzzle is that these are "private" or subjective phenomena: Only I can know what my sensation of blue looks like or what my pains feel like (or that I am in pain). You cannot know what my sensation of blue is like or what my pain feels like, or know that I have any blue-sensation or that I am in pain. You can, perhaps, feel a pain that "is like" my pain—though how would you (or anyone, for that matter, including me) really know that it "is like" mine, since you can only feel your own (cf. Smith's problem about knowing whether our models match the world, §2.7.1). In any case, your pain is not *my* pain. You can, perhaps, determine that I am in pain—but only on the basis of my publically observable physical behavior, and that, of course, could be mere show or—more radically—be "real" pain behavior unaccompanied by any qualitative painful sensation. So, qualia are private, hence "mental" (according to a well-accepted tradition). Hence, they ought to be explainable functionally or as part of the Mind Abstraction. Yet functionalism seems incapable of explaining them, or so the puzzle goes.

The way out, I propose, is *to view qualia as dependent on implementation side-effects*. This does not resolve the puzzle by itself, however, for we still have to account for its privacy.

Let's begin with the problem of "absent qualia": the possibility that, for example, I feel no pain in circumstances in which others do, yet I am not oblivious to the pain stimulus—I behave appropriately. Thus, an experimenter sticks pins in my right hand and in yours. We both wince, withdraw our hands, perhaps cry out; we both say that the pin-pricking hurts, and we complain of residual soreness over the next several hours. Yet you *feel pain* and I don't. Or so we suppose for the sake of argument. The questions are: (1) Is this possible? (2) Am I any "less" of a mental agent because of my lack of feeling? The issue is sharpened when I am replaced by a computer or, better, an android: Does the android feel pain? We suppose not. But why? The central issue here is one of subjectivity, the same issue that is at the heart of the Chinese Room Argument: Does an entity that passes a Turing-like test—in this case, one for pain/pain-behavior—"really" have the phenomenon being tested for? And, if *not*, does that mean, despite its behavioral indistinguishability from a human that *does* have the phenomenon, that it is only "going through the motions" and not "really" feeling, using natural language, or thinking?

I have mixed feelings about this (if you'll excuse the pun). On the one hand, I want to say that insofar as having—or lacking—the private sensation has *no* behavioral consequences (not even to my being able to describe my pain-sensation in exquisite and poetic detail—whether I have it or not), then it is *not* part of the Mind Abstraction. If I *do* feel pain, then my sensation must be due to my body—it is an implementation side-effect. I can, of course, perceive the pain sensation. Moreover, it *is* possible that the Mind Abstraction can deal with this despite the fact (if fact it be) that, despite the privacy, it is not a mental phenomenon: For the Mind Abstraction will have, let's say, a variable or data structure of some sort whose value *would* be the sensation if I *had* a sensation and whose value is unassigned otherwise. The assignment of a value to this variable or data structure is input from my body. That is how it is implementation dependent.

On the other hand, I think it is plausible that there are never any absent qualia. Take pain, and consider the following computational implementation of it suggested by Stuart C. Shapiro (in conversation): Imagine a computer terminal with a pressure-sensitive device hooked up to the central processing unit in a certain way that I'll specify in a moment. (All of this ought, by the way, to be able to be done with current technology.) Program the computer with a *very* user-friendly operating system that allows the following sort of interaction (comments in parentheses; cf. Figure 7.7):

(User logs in, as, say "rapaport")

System: Hi there, Bill! How are you? What can I do for you today?
(Assume that this only occurs at the first login; the operating system, assume, is capable of some limited, but reasonable, natural-language conversation.)

User: I'd like to finish typing the paper I was working on yesterday—file "book.30sep92".

System: No problem; here it is!
(The file is opened. The user edits the file, closes it, and then hits the terminal sharply on the pressure-sensitive device. Assume that this device is wired to the computer in such a way that any sharp blow sends a signal to the central processing unit that causes the operating system to switch from *very*-user-friendly mode to "normal" mode.)

System: File "book.30sep92" modified and closed. Next command:

User: I'd like to read my mail, please.
(System runs mail program without comment. (User exits mail program.)

System: Next command:
(User logs off; logging off in the context of having struck the pressure-sensitive device causes the operating system to switch to yet another mode. The next day, User logs in ...)

System: Rapaport. Oh yeah; I remember you. You hit me yesterday. That hurt!

Now, what's going on here? We have a computer with an artificial-intelligence operating system that is exhibiting pain behavior. Modulo the differences between the computer and a human, and the limitations of the natural-language interface, behaviorally (or, from the intentional stance) it is reasonable to infer (or assume) that the computer was in pain when I hit it. But did it *feel* pain? How do *humans* feel pain? We feel pain when certain neurons are stimulated and certain signals are sent to the brain. Now, in our computer, certain wires connecting the pressure-sensitive device with the central processing unit are "stimulated" and certain signals are sent to the central processing unit. Where's the difference between human and computer? Perhaps the difference is

"Now this one's really user-friendly."

Figure 7.7:

that, for humans, there is a "pain-sensing" neuron in the brain that is stimulated when a human is hurt. It gets its input from the pain neurons (C-fibers, or whatever), which also send their input to certain motor neurons that result in typical pain behavior (or perhaps the pain-sensing neuron sends its output to the motor neurons). Fine; build a similar such device into the central processing unit and operating system. The cases are parallel. There is a quale in both cases. Ah, but what does the computer's pain *feel like*, you ask? I don't know. Do you know what *my* pain *feels like*? We'll come back to this in a moment. My point, for now, is that pain qualia can *and will* arise whenever there is pain-behavior, and the same holds, *mutatis mutandis*, for any qualia.

Consider, next, the problem of "inverted" qualia (or, for that matter, "differential" qualia): The general problem of accounting for the particular "feel" of a qualitative experience, assuming the *presence* of qualia: Does your pain feel like mine? Does your sensation of blue look like mine? In the most perverse case—the inverted spectrum case—your sensation of blue is just like my sensation of yellow, and vice versa, all across the spectrum (possibly excepting a fixed point?); or, in the inverted-pain case, your feeling of pain is just like my feeling of pleasure, and vice versa. Can this be? How? Well, first, it seems plausible that something like this, if not quite so extreme, *can* be. There are the experiments with inverting lenses **[ref; check Cole 1990?]**, in which the subject becomes acclimated to seeing the world upside down—behavioral indistinguishability with distinct qualia. There appears to be no reason in principle not to be able to adapt this to inverted spectra (Cole 1990). And many of us can experience a similar phenomenon by closing one eye: In my own case, at least, colors appear distinctly different to each of my eyes; by crossing my eyes so as to produce a double image, I can even compare the differences in color.

Again, I suggest, this is merely an implementation-dependent side effect. Rather than speculating on how the brain might be wired, let's again consider a computer example. Consider two computer programs with the same input–output behavior, written in Pascal using stacks. Suppose that one of them implements the Stack abstract data type as an n-element array $A[0], \ldots, A[n-1]$ with $\texttt{top} = A[0]$, while the other implements it as an n-element array $A[0], \ldots, A[n-1]$ with $\texttt{top} = A[n-1]$. The internal mechanisms—the *implementations* of the stacks—are "inverted" with respect to each other, yet this is behaviorally undetectable and irrelevant. Granted, here there is no issue of "qualitative feel", perhaps. Yet the point is that the differences—and there clearly are differences, although not input–output ones—are implementation dependent. The analogue of qualia are implementation-dependent side effects. Similarly for pain: The *sensation* or *feeling* of pain, in humans, might be something that your body has (or does, or undergoes) when, for instance, you step on a tack. But that it feels the way it does is an epiphenomenon (so to speak) *of the body*. Were the same mind implemented in a different body (as in Leiber's *Beyond Rejection*), perhaps the feeling would be different (or absent).

Are qualia "mental" phenomena? They are private, yet (I hold) they are implementation dependent. Does that mean that functionalism (or strong artificial intelligence) fails to "model" some mental phenomena? That's certainly one interpretation, one move that can be made in the philosophical game. Or does it mean that what it fails to model (pain, spectra inversion) isn't mental? That is, of course, another equally plausible interpretation, another move that's open, *unless* one defines the mental in terms of what is "private" (i.e., not publicly accessible). Yet another option is that *some* of what we call 'mental' is body (or implementation) dependent, though this is not available for those who define bodily phenomena in "public" terms.

The position I find congenial is to make the "syntax" "complete". Recall my suggestion that implementation side-effects were due to situations where the semantic domain exceeded the syntactic domain. In such cases, we can *extend* the syntactic domain to make it match the semantic domain (cf. Rapaport 1981). Although any Mind Abstraction may be incomplete in *this* sense of having implementation side-effects, the *fact* of having such implementation side-effects can be made part of the Abstraction, as indicated earlier with my discussion of variables whose values are assigned externally. In this way, to paraphrase Tolstoy, every cognitive agent will "feel pain", but everyone's pain will "feel" different.

The random digits "sputtered" by a faulty chip that were interpreted as enemy missiles were also implementation side-effects—(physical) implementation details that yielded or gave rise to "mental" behavior: The computer interpreted certain physical configurations as meaning something (cf. Tenenbaum & Augenstein 1981)—it "felt" them in a certain way, so to speak. A feeling of pain is the mind's *perception* of a physical event. Thus, qualia can be thought of as the locus of "interaction" of mind and body, of Abstraction and Implementation.

It is not, therefore, unreasonable that qualia would be physical, yet "private" in virtue of being part of the Abstraction. The actual "feeling" belongs, and only belongs, to the implementation. Consider *Hamlet's* sadness (at, say, his killing of Polonius) as opposed to *Olivier's* sadness (at, say, learning of the death of a good friend) and as opposed to *Olivier-qua-Hamlet's* sadness. In the Method School of acting, Olivier-qua-Hamlet's sadness would be an implementation of Hamlet's sadness in the medium of Olivier's sadness. This is to be distinguished from Olivier merely "acting" sad (perhaps a case of absent qualia?).

The privacy of qualia just *is* its subjectivity. Compare the following three experiences:

Figure 7.8:

Seeing in a mirror something on your eyelash; seeing it (out of the corner of your eye) directly on your eyelash; feeling it on your eyelash. These are three distinct and different sensations, only one of which (the mirror case) is "objective" or external—from the third-person point of view.

Is there any *more* or *other* difference between these? I think not. Pain, etc., are just the way things are perceived in certain circumstances, some of which cannot be experienced by anyone other than the subject. "That certain feeling" *ought* to be private, because it's due to the *experiencer's* implementing medium, not anyone else's.

7.6.4 The Real Thing.

One fancy rephrasing of the question whether machines can think is this: Is a computer "simulation" of a mind "really" a mind? Compare this, for the moment, to another question (cf. Dennett 1978, Hofstadter 1981): Is a computer simulation of a hurricane "really" a hurricane? To answer questions like this, we need to spell out what a "simulation" is, what a hurricane (or a mind) is, and what it means to say that something "really is" an X.

The first observation I would like to make in this regard is that experience with an *implementation* of some Abstraction can change our understanding of the nature of the *Abstraction*. Can "straight lines" be implemented in a non-Euclidean geometry? The answer is 'Yes', but they aren't "straight" anymore; they can only be implemented as *geodesics*: shortest distances between two points. So, on a sphere, the implementation of straight line is a great circle. Similarly, consider the implementation by airplanes of flying: Airplanes fly, but not the exact way birds do (e.g., although their wings might have more or less the same shape, planes don't flap them). Planes fly only (or at least?) in the sense of moving through the air without touching the ground. No doubt that needs to be refined, so as to rule out long jumps (but mightn't a *very* long jump *be* flying?). Yet another refinement might replace the reference to air with a general term for a fluid medium: It has been suggested that the knowledge-representation community's favorite flightless bird—the penguin—does indeed fly … in water (Ackerman 1989: 45–46). The point is, as we saw earlier

(§7.6.1), that when an Abstraction is implemented in different media, there will be implementation-dependent differences, yet there will be some common essence to both, in virtue of which they can be said to be the same. Thus, what is "really important" about straight lines is that they are geodesics; that geodesics are "straight" in Euclidean space is an implementation-dependent side effect—an "accidental" property, if you wish.

Edsgar W. Dijkstra has been quoted as saying that "the question of whether a computer can think is no more interesting than the question of whether a submarine can swim" (cf. Mike Gobbi, posting on the sci.psychology newsgroup, 20 September 1994). Assertions of equivalence such as this are notoriously ambiguous. Does Dijkstra think that it's obvious that submarines *do* swim (and therefore that computers *do* think)? Or that it's obvious that they don't? Or that it's merely a question of whether we'll extend the meaning of 'swim' to cover whatever it is that submarines do? Suppose the latter. What is it that submarines do? They move in the water. But that's what swimming is,[11] though perhaps before the advent of submarines we thought that swimming *had* to be done by animals: Do fish swim? Surely. Do people? Perhaps only by extending the term. Extending the meaning of a term occurs when we realize or decide that a property that we thought was essential isn't. This goes a long way toward explaining the unease people feel when they're told that computers can think.

So, is this extension of terms such as 'fly' to planes, 'swim' to submarines, and 'think' to computers "merely" a metaphorical extension? It may be metaphorical, but it is not "mere":

> **Eus[ebius]:** ... I do wish you would stop using terms borrowed from human behavior [to describe monkey behavior]! You're being anthropocentric!
>
> **Soc[rates]:** Well, monkeys are anthropoids. Besides, do you want me to make up a new word for a phenomenon for every species that shows it? Should geneticists stop talking about inheritance because that term was borrowed from economics? (Altmann 1989: 260.)

There are two points: First, *refraining* from such extensions, metaphorical or otherwise, would force us to miss important generalizations. Second, as Lakoff and Johnson (1980) have shown us, metaphor is an unavoidable and central feature of our language and thought.

What we do have to be careful about is mixing our metaphors. That is, an implementation must be complete unto itself; we must not import or apply features from one implementation of an Abstraction to another implementation of the same Abstraction. Thus, to take the classic case, it is of course not true that computer-*simulated* hurricanes get *real* people wet. But they *do* get *simulated* people *simulatedly* wet (Hofstadter 1981; cf. Rapaport 1988, Shapiro & Rapaport 1991). "Obviously, a computer simulation of a stomach would only digest simulated food" (Johnson 1990: 46). And a "simulated engine wouldn't generate any 'here in the world outside the computer' power—but if you put it in a suitably simulated car, and engage the suitably simulated clutch, it will just fine drive down the simulated road" (Minsky 1991). In each of these cases, we do have "the real thing": A simulation of digestion *is* digestion, a simulated hurricane *is* a hurricane. More accurately, I propose, a computer simulation of human digestion is an *implementation* of the Digestion Abstraction, as is human digestion itself. The latter may be more familiar, more prototypical (cf. Rosch 1978), but both, just as Dijkstra observed of swimming, are really digestion.

[11] Unless, of course, swimming is flying in water!

(apparent?) input	→	P88 program	→	(apparent?) output
		↓		
		P88 machine-language program		
		↓		
		Pascal program		
		↓		
(actual?) input	→	Macintosh machine-language program	→	(actual?) output

Figure 7.9: Hierarchy of virtual-machine levels.

The difference between a "simulation" of flying or of a hurricane and what we normally think of as "real" flying or a "real" hurricane is that the former "are one step removed from reality ... [because they use] symbolic parameter values that represent physical behavior" (Johan Lammens, personal communication, 17 August 1990). I'm not sure about the latter, but I agree with the former: Computer simulations are not part of the "real" world (except, of course, in the sense in which *everything* is part of the real world). They exist in their own simulated world, and we must be careful about "transworld" attributions. Although a simulated hurricane will not get *us* wet—because that would require a "transworld" causal relation of a kind that does not exist—the simulated hurricane must have some of the "same" (or analogous) cause–effect relationships with denizens of *its* computer universe (e.g., getting simulated people simulatedly wet) in order for it to count as a simulation—in order for both it and 1992's Hurricane Andrew to be *implementations* of the Hurricane Abstraction.

There is, however, an important family of exceptions to this principle of segregation. Computer simulations of semantic or information-processing systems are not only implementations of them, but can interact with other such implementations. In other words, if they were simulated hurricanes, they *could* get *us* wet.

A clear example of this is the computer simulation of computation itself. In one of my introductory computer science courses, I occasionally use a piece of software called the "P88 Assembly-Language Simulator". "P88" is (a fragment of) an assembly language for a hypothetical machine (Biermann 1990). (We can ignore for now whether it really *is* an (incomplete) assembly language *even if* "just" a toy one, since that is irrelevant to the point I want to make.) The P88 Assembly-Language Simulator is a Pascal program (actually, a ThinkPascal program). As such, it must be compiled into the machine language for the Macintosh computer on which it runs. My students and I can write P88 programs and "assemble" them into (a simulation of) a P88 machine-language program, which, in turn, is interpreted by Pascal as a certain Pascal program, which, in its own turn, is compiled into a Macintosh machine-language program, and executed. The levels are shown in Figure 7.9.

Suppose, now, that I write a program *in P88 assembly language* that takes two integers as input and returns their sum as output. When I cause this P88 program to be executed, a prompt appears on the screen, I input an integer, another prompt appears, another integer is input, and their sum is printed on the screen as output. Question: Was this a P88 computation? Answer (as usual): Yes and no.

Yes: It was, because the *algorithm* that computationally caused the sum of the two inputs to be output was a P88 algorithm that used data structures and instructions from the P88 language.

In other words, the two integers that were the input to the P88 program were input *to that program*, and their sum was output *by that program*.

No: It was not a P88 computation, because the P88 algorithm was executed by having a *Pascal* program perform *Pascal* computations that used data structures and instructions from the Pascal language. So, was it, then, a *Pascal* computation? Well, in some sense not really, because the Pascal algorithm was executed by having a Macintosh machine-language program perform *Macintosh machine-language* computations that used data structures and instructions from the Macintosh machine language. (Curiously, these are more like the data structures and instructions from P88 than from Pascal, but they were not executing or simulating P88 instructions or using P88 data directly.) At bottom, then, only a Macintosh machine-language program was really being executed and really computing the sum of two integers. In other words, the two integers that apparently were the input to the P88 program were actually input to the Macintosh machine-language program, and their sum was actually output by *that* program.

Yet I can, and do, use the P88 Assembly-Language Simulator to compute the sum of two integers. In other words, a computation by the P88 program was simulated by a Macintosh machine-language computation, but there *was* a computation nonetheless. The simulated computation *was* a computation. Moreover, it was a computation in two senses (once could say that there were two (simultaneous) computations (with the same input and output)): Ignoring *how* the P88 program was implemented, a P88 computation did yield the sum of the two inputs. And, ignoring the fact that it was simulating a Pascal simulation of a P88 program, the Macintosh machine-language computation also yielded the sum of the two inputs. It is important, I think, to note that the Macintosh machine-language program did so in a roundabout way: It was *not* the simplest possible such program to output the sum of two inputs, because it did that *not* simply by performing an addition, but by doing a number of other ("bookkeeping") operations that simulated a Pascal program simulating a P88 program.

As in Figure 7.8, mental "fingers" are an implementation (a semantic correlate) of actual fingers, and can serve (some of) the same purposes. Elizabeth (the child in the cartoon) is not using her actual fingers, and she *is* doing the work in her head—by using "head-fingers". Imagination and mental imagery can serve as a substitute for actual experience—one can solve problems by manipulating either the actual objects or models of them (see Figure 7.10).

Let's consider some examples, some of which we've seen in other contexts. A photograph of a map (cf. §2.7.3) can be used to find out where some city is, even though that wasn't the purpose or intent of the map in the ad. Copying information (*sic*) from a book (by Xerox or by hand) and then using that copied information (those copied sentences) rather than the original source is done all the time. We don't think twice about it or say that it's not "really" information. Here, unlike the previous example, there *is intentionality*.

Why is it information? Because of its syntax and the reader's interpretation of it. But it *carries* the information whether or not a reader interprets it. There is a possible problem: The information carried could be differently interpreted by another reader. But what's invariant is the syntactic structure. In any case, multiple interpretations are all equally good interpretations. Ah, but *is* the syntactic structure invariant? Examples such as the string 'NOWHERE', which could be analyzed as 'NOW HERE' or as 'NO WHERE', or weakly equivalent grammars suggest that the larger the context, the more aid there is in determining the syntactic structure.

What about computer simulations of minds? "The difference between a symbolic airplane

Figure 7.10:

simulator and a symbolic intelligence simulator is that the former models a physical system through the intermediary representation of parameter values, while the latter models behavior by behaving" (Johan Lammens, personal communication, 17 August 1990). Here's the insight: *We* don't interact with simulated hurricanes, so they don't get *us* wet. (As we've seen, they *do* get simulated people simulatedly wet—that's where the "internal representations" come in.) But we *do* interact with simulated minds. Now, how is that possible? *Is* it possible?

Do we thus interact, or do we only seem to? Having a conversation on some topic with a Turing-Tested simulated mind *is* having a conversation on that topic and *not* merely simulating having the conversation. For the latter is what you would do in a play. "When you get out of a TT session, something has changed: you have talked to a system about something, and most likely that has affected some of your own thoughts and beliefs" (Johan Lammens, personal communication, 17 August 1990). Of course, being in a play can do that, too, just as *reading* the play could. But in the Turing Test case, it's a dynamic, changing conversation.

So, how is the interaction possible? Because both systems deal with information, albeit implemented differently. But the implementations are "transparent" (as in the game of chess played with Staunton pieces; cf. §3.2.1).[12] One must, of course, be careful to distinguish—if possible—between information that only concerns the simulated world (or, for that matter, the real world) and information that can transcend the boundary. Smith cites an example of a training tape that was interpreted to be a real Soviet attack (1985: 00). **[???]** The very fact that this can happen shows that some "simulations" are indistinguishable from the real thing. Indistinguishable in what sense? In the Fregean *Sinn*-sense or Meinongian-object sense, not the Fregean *Bedeutung*-sense or the sense of actual objects: In the one case, there is a real-world referent (or Sein-correlate), but not in the other. At the level of *Sinn* or Meinongian object (or mental representation), all things are on a par. If we can't determine that they differ referentially, the default assumption should be that they don't. And even if we *can* determine it, it might not matter. (For more discussion of this, see Rapaport 1991a.)

So there is a difference between the digestion and hurricane cases on the one hand and the natural-language and mind–brain cases on the other: "... brains, unlike stomachs, are information-

[12] This replies to Jahren's objection in his footnote 1, p. 326. **SPELL THIS OUT**

processors. And if one information processor were made to simulate another information processor, it is hard to see how one and not the other could be said to think" (Johnson 1990: 46). That is, the difference is that it *is* the same stuff involved in the brain and computer cases: "Simulated thoughts and real thoughts are made of the same stuff: information" (Johnson 1990: 46). Well—not quite: Information is abstract; simulated and real thoughts are different *implementations* of the same *Abstraction*. (One might, however, want to say that they deal with the Abstraction *directly*, via "transparent" media.)

My claim, then, is that simulated mentality (or cognition, or intelligence) *is* mentality (or cognition, or intelligence). Recall the mental imagery debate, in which Stephen M. Kosslyn (1981; Kosslyn et al. 1981) argues that one really scans a mental (i.e., simulated) image, whereas Zenon Pylyshyn (1981) argues that one *pretends* to scan (*simulates* scanning) a real image. Or compare Searle's claim (1979) that in fictional language, one *pretends* to assert (one *simulates* asserting) rather than *really* asserting a pretended (or simulated) utterance—an utterance in a pretend-world.[13] In both of these cases, I side with the "really" people: Rather than saying that a computer *simulates* understanding something real, I would say that it *really* understands something simulated—and that in many cases, the simulated thing that it really understands is itself the real thing (internally represented).[14]

7.7 FROM MULTIPLE REALIZABILITY TO PANPSYCHISM.

Given the Principle of Multiple Realizability—the apparently obvious claim that there can be more than one implementation of an Abstraction, more than one model of a theory—an argument can be constructed for a variety of panpsychism (the view that everything is a mind). The argument, in its bare outlines, is this:

1. There is multiple realizability (of computational processes).
2. ∴ There is universal realizability (of computational processes) (by an argument of Searle's).
3. ∴ Anything can be a model of anything (else) (from (2), or by an argument of Wartofsky's).
4. ∴ Anything can be a model of a mind (from (3) by universal instantiation, or by an argument of Dipert's).
5. ∴ Anything can be a mind (by the argument of §7.6.4 that *models* of minds *are* minds).

Actually, as will be seen, there are several different arguments in this vicinity. Let's begin with Searle's argument from (1) to (2).

[13] Note the deictic shift; cf. Galbraith**?**, Segal**????**

[14] B. H. Webb (1991: 247) argues "that it is possible for a simulation to be a replication if the device used can not only represent but also instantiate the same capacities as the system." This seems congenial to my claims.

7.7.1 Multiple Realizability Implies Universal Realizability.

In "Is the Brain a Digital Computer?" (1990), Searle is concerned with the multiple realizability of computational processes. Hence, on the assumption that mental processes are computational, he is concerned with the multiple realizability of mental processes. A "disastrous" consequence of multiple realizability is that it

> would seem to imply universal realizability. If computation is defined in terms of the assignment of syntax, then everything would be a digital computer, because any object whatever could have syntactical ascriptions made to it. You could describe anything in terms of [**???**] 0's and 1's. (Searle 1990: 26.)

What evidence does Searle have for this claim that any (physical) object can be described computationally? And why is it disastrous? The latter question is easier to answer: According to Searle, universal realizability doesn't tell us what's *special* about the brain as opposed to other, less interesting, computational systems, such as the "stomach, liver, heart, solar system, and the state of Kansas" (p. 26). Perhaps—we'll come back to this (§7.7.3).

Searle claims that "For any object there is some description of that object such that under that description the object is a digital computer" (p. 26). This seems too strong. For one thing, there are certainly things that are non-computational in the strong sense of the Church–Turing Thesis. For another, merely assigning 0s and 1s to give an encoding (of, say, the atomic structure) of a physical object doesn't make it computational. To be computational, analogues are needed of Turing-machine instructions, control structures, states, the input–output tape, etc. At the very least, to be computational, an item must compute some function. So, what function does my pen compute? Well, I suppose it could be argued that it computes the constant function (or perhaps the identity function, or perhaps it loops forever—i.e., is undefined on all input). But that is trivial. On the other hand, consider the string of 0s and 1s that, according to Searle, encodes my pen. That's the Gödel number of some program (no doubt a trivial one, but who knows?). Does that make my pen an *implementation* (a model) of that program? No; it is an interesting correspondence, but not an implementation, because the interpretations of the 0s and 1s for the description of the pen are not those required for the program.

Smith (1982) has made similar observations. He describes a "computer" that "calculates oriental trajectories", which is, in fact, a car that drives west,[15] and he asks why this *isn't* a computer (p. 2). He notes that it does share two important features with computers, the second of which is close to Searle's claim: First, the oriental**???**-trajectory calculator *is* equivalent to a Turing machine, so that's not why it isn't a computer. But *why* is it equivalent to a Turing machine? Perhaps because it is input–output equivalent to a Turing machine that calculates oriental**???** trajectories? If so, then perhaps mere input–output equivalence is not sufficient: As we saw in §9.5, that a *function* is comput*able* merely means that there is a Turing machine that computes it (i.e., that has the same input–output behavior), but that does not mean that a device with that input–output behavior is *itself* a computer (that it *computes* its output from its input)—it could be a mysterious oracle. Now, the case of the car is not quite the case of such an oracle. The car, after all, has parts whose function and behavior contribute to the car's overall behavior (its output,

[15] Shouldn't that be *east*? Perhaps 'oriental' is a typographical error for 'orbital'? Or perhaps he means 'oriental' as in "pertaining to orienting oneself".

if you will). So why does Smith say it's not a computer? Because there are no *symbols* that "act as causal ingredients in producing an overall behavior" (p. 2)—symbols in the sense of markers of a formal syntactic system:

> In describing how a car works, ... the story is not computational, because the salient explanations are given in terms of mechanics—forces and torques ... and so forth. These are not *interpreted* notions; we don't posit a semantical interpretation function in order to make sense of the car's suspension. (Smith 1982: 3.)

I dispute that. First, there is a mapping between the physical parts and actions of the car and terms from physics (i.e., physical theory). Second, there is a mapping between (at least) terms from physics and my concepts. So we *do* interpret the car's parts and actions.

Here is Smith's response to the claim

> ... that we 'interpret' steering wheels as mechanisms for getting cars to go around corners—... this is a broader notion of 'interpret' than I intend. I mean to refer to something like the relationship that holds between pieces of language, and situations in the world that those pieces of language are about. (Smith 1982: 4.)

But that distinction is one that can't be drawn (or so I tried to argue in Chapter 2)—*both* are interpretations. The threat of universal (or near-universal) realizability is expressed by Smith thus:

> ... the 'received' theory of computation—the theory of effective computability that traffics in recursive functions, Turing machines, Church's Thesis, and the rest ... does not intrinsically identify the class of artefacts that computer science studies. ... [I]t is too broad, in that it includes far more devices within its scope (like chairs and Rubik cubes) than present experts would call computers. The problem stems from the fact that Turing equivalence (i.e., computing the same function) is a *weak, behavioral* metric, and we are interested in a theory that enables us to define *strong, constitutional* concepts. (Smith 1982: 5.)

I'm willing to accept Rubik's cube, however. The difficulty is that "the class of artefacts that computer science studies" is an *intended interpretation* that the theory of computation just won't let us get our hands on, any more (but, equally, no less) than Peano's axioms let us get our hands on the natural numbers. If chairs are included (as Dipert argues; see below), so be it. Even if we strengthen the theory to talk about *algorithmic* equivalence, and not mere input–output equivalence, we'll still get multiple, hence unwanted—or, better, *unexpected*—realizations.

7.7.2 Everything Models Anything.

If there is universal realizability, then anything can be a model of anything (else). This is because, from the assumptions that, given an arbitrary Abstraction, everything can be an implementation (or realization) of it, it follows by universal generalization that everything can be an implementation (or realization, or model) of anything. Alternatively, as we saw in our discussion of Wartofsky in Section 2.6.2, step (3) of the argument for panpsychism follows from the assumption that everything shares at least one property (and perhaps infinitely many) with everything (else).

7.7.3 Everything Models Mentality.

Clearly, if anything can be a model of anything (else), then anything can be a model of a mind. An argument explicitly for this consequent has recently been offered by Randall R. Dipert (1990).[16]

Dipert begins by reminding us that David Hilbert's philosophy of formalism took numbers in "purely structural, formal terms ... [C]hairs, beer mugs, or whatever could just as well represent/exemplify numbers (under the right interpretation) as do numerals or our thoughts of numbers" (p. 6).[17] Similarly, "programs, together with their hardware implementation ... may not look much like more usual [i.e., biological] embodiments of minds" (p. 6), but they could be, in the same way that chairs can be embodiments of numbers. However—or so Dipert observes—not even adherents of so-called "strong AI" would take *chairs* as embodying minds, because "brains are much more complex than chairs, and so chairs and tables lack some of the structural features of mental properties" (p. 6). However, I am an adherent of "strong AI" and am as willing to accept the claim about chairs as I am about the standard water-pipes-and-valves model, which I am quite willing to do. Dipert (along with Searle and Smith, evidently) thinks this is problematic. Here's the argument that shows why:

(P1) Ordinary, middle-sized objects at room temperature (let's call them OMSORTs, for short)—e.g., chairs, coffee mugs, baseballss (and, presumably, brains)—are highly complex, dynamic entities (pp. 7–8).

(P2) Suppose there is a good cognitive science theory T of the sufficient conditions for "cognition and other mental processes" (p. 8).

(P3) Suppose there is an AI system C that implements T (p. 9).

(C1) $\therefore$ By formalism, C has cognition (p. 9).

(P4) For every OMSORT O, there is an interpretation such that O exemplifies T (although we might not be able to exhibit the interpretation that does the job) (pp. 9, 11).

(C2) $\therefore$ By formalism, O has cognition (pp. 9–10).

> These reflections should also make us resist our initial temptation to say that exemplifying some humanly-graspable ... set of properties [is] *sufficient* for having mental properties—unless we are willing to say, with Leibniz, that everything is a mind. (Dipert 1990: 11.)

Thus,

(P5) Conclusion (C2) is absurd or uninteresting.

(C3) $\therefore$ Conclusion (C1) is absurd or uninteresting.

[16] Dipert himself is sympathetic to the conclusion, even though he is playing devil's advocate in criticizing it; cf. Dipert 1990: 20n16.

[17] Cf. Hilbert's *Gesammelte Abhandlungen*, vol. 3, p. 403, as cited in Coffa 1991: 135.

Now, one difference between the arbitrary OMSORT O and the AI system C is that for C we *do* know what the interpretation function is: We can *understand* how and why C behaves as it does; we can interpret C's behavior as a mind. We accept it as such (this is what we do in our everyday solution of the problem of other minds).

Note, too, that some OMSORTs that we *might* very well be willing to accept as implementations of minds—namely, connectionist implementations that haven't been "properly treated" (Smolensky 1988)—are such that we might very well *not* understand them (what, e.g., do the connection weights "mean"?).

What's wrong with Leibniz's position? Mainly that if everything is a mind, then we can't explain the difference between a *human* mind and a coffee mug. In this regard, we might be no more worse off than a topologist who, as the joke has it, can't distinguish a doughnut from a coffee mug, since both are toruses. To say that there is a way to view doughnuts and coffee mugs such that they are alike is *not* to say that there is no way in which they differ. Similarly, if computational cognitive science tells us that, from a certain perspective, brains and mugs are alike, that's not to say that, from some other perspective, they're not. We *can* explain the difference between a mind and a mug: The mug mind can't communicate with us and therefore is irrelevant. That is, the mug *qua* mind can't communicate; the mug *qua* coffee-holder is a perfectly functional device. Brains *can* communicate, but they can't hold coffee—so we don't use them for that purpose. Not to be too macabre, we *could* use a brain as a paperweight, I suppose; a Martian (or a Black Cloud) might, and might never realize that its paperweight implemented a mind, any more than we realize that a coffee mug might. As a further analogy, compare a high-level program (e.g., to compute greatest common divisors, or even an AI program) with a machine-language version in the same way we have been comparing brains with mugs. The latter might not *look* like the former, but under the right circumstances it might very well behave like the former.

What, then, is the import of Dipert's argument? Is it merely that, for any theory, there are infinitely many models, many of which are non-isomorphic (cf. Mac Lane 1981: 467) and many of which are not the intended model(s) and are such that we did not antecedently take them to be models? If so, so what? Sure, there *have* to be unintended models, and there is no way to pick out or mark or identify the intended ones; that's one of the main lessons of the theory–model relationship. What we learn from the existence of unintended models (assuming that we are completely satisfied with the *theory* of which they are models) is that some things have properties and features that we didn't expect them to have.

Of course, Dipert's transcendental argument merely shows that it's highly likely that OMSORTs can be taken as (models of) minds, not that they *are*. Two highly complex things, just because of their high complexity, need not be models of each other. (Two highly complex patterns need not be matchable.)

On the other hand, Dipert's claim might be the weaker one that (it is highly likely that) there are *some* OMSORTs that model the cognitive theory T but that we would not antecedently have taken as an intended model. I think we can only bite the bullet on this. But perhaps it can be made palatable: Suppose the OMSORT is a (particular) baseball. Imagine complex dynamic *processes* "within" the baseball, presumably at the subatomic level, that model the mental processes. What, for instance, might correspond to perception? (Does the baseball "see"?) Perhaps nothing so corresponds in the sense of external causes of internal processes, but there might be internal processes that, in a methodologically solipsistic fashion, correspond to (or model) perception. Or perhaps there *are* such external causes, but they need not be actual events as *we* characterize (or see) them. The world that the baseball-mind "perceives" might (indeed, probably would) have different categories than the human mind or the AI mind (as Kant told us long ago; cf., too, Winston's Problem (§3.2.2.2.4) and Kirsh 1991: 12).

7.8 SUMMARY.

In this chapter, we have investigated the mind-brain relationship as a case of implementation. I have proposed that implementation is best understood as semantic interpretation (rather than as individuation, instantiation, reduction, or supervenience). It is a relationship between an Abstraction (a generalization of the notion of an abstract data type) and an implementing medium. This relationship can be found in music and language, as well as in the theory of abstract data types. In general, something is an implementation of an Abstraction in an implementing medium. Mind is an Abstraction that can be implemented in brains as well as in computers. Implementations, however, have implementation-dependent details that give rise to qualia—implementation side-effects.

If Mind can be implemented in a computer, could a computer that implemented a natural-language–understanding program really understand language? I, of course, would say 'Yes'. Searle, of course, says 'No'. So let's return to Searle's Chinese room, in the light of our present conclusions.

Chapter 8

RETURN TO THE CHINESE ROOM.

8.1 INTRODUCTION.

I began with the question of how we can have knowledge of the semantics of our language. This is the challenge posed by Searle in his Chinese-Room Argument: How could Searle-in-the-room come to have knowledge of the semantics of the (Chinese) squiggles? For Searle, it is the task of explaining how Searle-in-the-room could know what the symbols *are about*, not what their syntax is.

Could Searle-in-the-room come to know the syntax? Not, presumably, *just* by having the sort of program that Searle envisions (namely, a Schankian Script Applier Mechanism (SAM, Cullingford 1981). A syntax-learning program is necessary. There has certainly been a lot of work on the problem of learning syntax, so this is not science fiction. So let's assume, as we did in §2.8.2, that Searle-in-the-room has such a program. Given an understanding of the syntax, how much semantics can be learned? We saw an answer to this in Chapter 2.

8.2 WHAT IS IT LIKE TO BE IN A CHINESE ROOM?

It might prove useful, at the beginning, to consider a few *actual* situations that can give the reader some semblance of what it might be like to be the person in the Chinese Room.

8.2.1 The Japanese Room and Subjective Experience.

The first situation is the one discussed in Chapter 3: the paper about SNePS written in Japanese (Arahi & Momouchi 1990) and well illustrated with SNePS networks, each of which have *English* arc labels but *Japanese* node labels for the `LEX` nodes, as in Figure 3.8. This is, of course, a reasonable thing to do. After all, Shapiro and I have repeatedly written that the *arc* labels convey no information to the system; they serve as "punctuation" or structuring devices only. True, the "reader" of a network uses the arc labels to help understand the network, but that's akin to the use of self-commenting variable names in a high-level programming language: useful to an "external" reader but not essential to the interpreter. True, too, SNePS's natural-language module uses the arc labels to generate appropriate English expressions, but—again—the arc labels could equally well have been more-or-less arbitrary strings of characters—what counts is that the arcs are labeled, not what the labels are.

Furthermore, as Shapiro and I have repeatedly urged (cf., e.g., Rapaport 1988), the "*content*" of the networks is in the `LEX` nodes. So it is appropriate for the Japanese networks to have English (better: *non*-Japanese) *arc* labels and *Japanese* `LEX`-node labels. The networks are, thus, general …

… And incomprehensible! When I saw them, I felt like Searle-in-the-room. I can manipulate the networks, but I don't know what they mean. What's missing is a link—a correspondence—between the `LEX` nodes and *my* `LEX`-node analogues *or else* between the `LEX` nodes and things in the world. The latter, though, is impossible; as we've seen, at best that would be *representations* in my mind *of* things in the world. So this reduces to the first alternative.

And it is that first alternative that captures, I think, Searle's frustration. The fallacy is that Searle would want to understand the network semantically, whereas the only option open to him is to understand it syntactically. Granted, there aren't enough samples of the Japanese networks for me to be able to make much sense out of them. But given enough networks and time, I would be able to.

But where, as Mary Galbraith has urged upon me (cf. Galbraith & Rapaport 1991), is the *experience* of understanding? Where is that feeling of "Aha! *Now* I understand"? The short answer is that the understanding is "in" the complex network. Now, on the face of it, that's not very satisfactory. If I can detect no understanding from a few networks, why should it arise from a few more? The long answer is that the more "interlocking" networks there are, and the more experience the person in the room has in manipulating them, the more understanding there will be. Part of this comes from "meta-understanding": "I understand what this means" is itself (an expression of) a piece of network that "refers" (in the manner of §3.2.2.1) to other pieces of the network.

8.2.2 The Library Room.

For a second situation, consider my understanding of Library of Congress catalog numbers: I don't know the rules of syntax or semantics for them, nor do I need to in order to use them "fluently" to find books. However, I have come to learn, inductively, certain rules of syntax (e.g., LOC-catalog-number ::= letter$_1$ + number + '.' + letter$_2$ + number + year) and of semantics (e.g., meaning(letter$_2$) = initial letter of author's last name, usually; meaning(year) = year of publication, usually; letter$_1$ categorizes certain books together; etc.). The more links I make with my knowledge of books, the more I know of the syntax and semantics (though it doesn't necessarily help me to communicate any better); and the more I know of the syntax and semantics, the more I understand of what I'm doing. Searle-in-the-Library-of-Congress-room would also come to have such understanding. Why shouldn't Searle-in-the-Chinese-room?

8.2.3 The Helen Keller Room.

But this is all fiction and speculation, the skeptic says. Has anyone ever been *in* a Chinese Room for a sufficiently long time? The answer, surprisingly, is 'Yes': Helen Keller has.

> The morning after my teacher came she ... gave me a doll. ... When I had played with it a little while, Miss Sullivan slowly spelled into my hand the word "d-o-l-l." I was at once interested in this finger play and tried to imitate it. When I finally succeeded in making the letters correctly I was flushed with childish pleasure and pride. Running downstairs to my mother I held up my hand and made the letters for doll. *I did not know that I was spelling a word or even that words existed; I was simply making my fingers go in monkey-like imitation.* (Keller 1905: 35; my italics.)

At the beginning of this passage, one expects that the antecedently played-with doll would be associated with the finger-spelled word 'd-o-l-l'. But as can be seen from Helen's later claim of ignorance ("I did not know ..."), her statement that she "made the letters *for doll*" (my italics) must be taken *de re* (they were letters-for-dolls), since, clearly, Helen did not know that she was "making ... *letters*" (my italics) or that they were "for doll". The last sentence is, I think, a significant passage. It is a wonderful description of pure syntax. Searle would be pleased. Annie Sullivan (Helen's teacher), on the other hand, no doubt would have had reason to believe that Helen *did* know what she was doing. Annie plays native-Chinese speaker to Helen's Searle-in-the-room.

The passage continues:

> In the days that followed I learned to spell in this uncomprehending way a great many words, among them *pin, hat, cup* and a few verbs like *sit, stand* and *walk*. But my teacher had been with me several weeks before I understood that everything has a name. (Keller 1905: 35.)

Again, these are *de re* descriptions of her own experiences, given long after the fact. She could experience external things and could experience meaningless finger manipulations, but she did not link them. Such linking between a word (a finger spelling) and an external object would have yielded semantic understanding. They would have played different roles: one the role of syntax, one that of semantics. One (the finger spellings) would not have been comprehended; the other (the physical objects) would have been familiar (they were, after all, part of the world she lived in every day). One would have been a name for a thing, the other a thing named.

Semantic understanding, as I have been at pains to show throughout the preceding chapters, would actually have come via Helen's linking of *internal* representations of *both* of those external experiences. And, as we know, she succeeded remarkably well. So, I suggest, would Searle-in-the-room.

At the end of this chapter and in the next, we'll return to Helen's story. What I hope to have done here is to provide some "intuition pumps" for the discussion to follow.

8.2.4 The Chinese High-Rise Apartment House.

Consider one final situation. My desk consists of pieces of wood, which consist of certain organic molecules, which consist of certain atoms, which consist of certain subatomic particles (electrons, protons, neutrons), which consist of certain quarks and leptons, etc. Similarly, Cassie is implemented in (a particular version of) SNePS, which is implemented in (a particular dialect of) Lisp, which is compiled into a particular machine language, which—when loaded—sets the values of certain hardware registers (i.e., "sets the switches"), which—when a certain signal is transmitted—causes the computer to (let's say) converse with me in English or Chinese.

The Chinese "room", then, is more like a multi-storied, high-rise apartment house. I'll refrain from pushing that analogy beyond its breaking point. What I am trying to remind you of is that there are *levels* of virtual machines—a hierarchy—the "top" one of which is the one that, apparently, is "really" Cassie and the "bottom" one of which is, apparently, "merely" a machine. (Recall the P88 example from the previous chapter.) The "virtual person" at the top may be distinct from any component of the hierarchical system. David Cole (1991) has argued thus; I have argued that if Cassie "resides" at the top level, she also "resides" at all lower levels (Rapaport 1990; for an overview of this debate, see Bringsjord 1992, Ch. 5). However, both of us agree, I think, that Searle-in-the-room does not understand Chinese *qua* Searle but only qua (a part of) one of the levels of the hierarchy. (For a nice attempt at making such a virtual person plausible, see Suits 1989.) John McCarthy has said something similar: "The Chinese Room Argument can be refuted in one sentence: 'Searle confuses the mental qualities of one computational process, himself for example, with those of another process that the first process might be interpreting, a process that understands Chinese, for example'" (1990).

Although the popular view of computers is that they are machines that "obey instructions", and the standard architecture for expert systems consists of an inference engine that *applies* rules stored in a knowledge base to situations in the world, computers don't "obey" and it is not the inference engine that is the expert. Patrick Hayes has said it best:

> The basic flaw in Searle's argument is a widely accepted misunderstanding about the nature of computers and computation: the idea that a computer is a mechanical slave that obeys orders. This popular metaphor suggests a major division between physical, causal hardware which acts, and formal symbolic software, which gets read. This distinction runs through much computing terminology, but one of the main conceptual insights of computer science is that it is of little really scientific importance. Computers running programs just aren't like the Chinese Room.
>
> Software is a series of patterns which, when placed in the proper places inside the machine, cause it to become a causally different device. Computer hardware is by itself an incomplete specification of a machine, which is completed—i.e. caused to quickly reshape its electronic functionality—by having electrical patterns moved within it. The hardware and the patterns together become a mechanism which behaves in the way specified by the program.
>
> This is not at all like the relationship between a reader obeying some instructions or following some rules. Unless, that is, he [sic] has somehow absorbed these instructions so completely that they have become part of him, become one of his skills. The man [sic] in Searle's room who has done this to his program now understands Chinese. (Hayes 1990.)

Just as we don't ask whether it's the human or his brain or her cortex (or whatever) that understands language, so we shouldn't ask whether it's Searle or his squiggle–English translation handbook (or whatever) that understands in the Chinese Room, or whether it's the central processing unit (or whatever) that understands in Cassie's case. This is, in part, the so-called "systems reply" (Searle 1980). It's the whole room, the whole person, the whole computer that understands.[1]

Using his *own* notions and subjective experience of understanding language, Searle cannot come to know whether or how a computer or—more to the point—even Searle-in-the-room understands Chinese. They are different kinds of experiences. Perhaps it is not unlike the situation of "neurologically impaired people such as autistic individuals [who] do have consciousness, but ... [one that] is different from ordinary people's and therefore results in a different experience of the world" (Lynne Hewitt, personal communication; cf. Winston's Problem). Hewitt goes on to say that she "doesn't believe that you can produce linguistic competence by providing more facts to the communicator: autistic people sometimes have an amazing ability to learn long stretches of language verbatim, while lacking the ability to understand why most ordinary people engage in ordinary conversation." It's the last that's important: The Chinese Room system wouldn't have linguistic competence unless it *understood why* it was conversing. Of course, its actual conversation could include plausible answers to such questions as "Why are you conversing?" without the computer *wanting* to converse. To get the latter, we would need, as noted in §1.2.4, a theory of when and how a computer could *initiate* conversation. One way is via *questions* that the computer raises while trying to analyze its data or understand the meaning of a new word (cf. Colby & Smith 1969; Ehrlich & Rapaport 1992, 1993, 1995; Ehrlich 1995).

8.3 SEARLE ON BRAINS AS COMPUTERS.

In his 1990 essay, "Is the Brain a Digital Computer?", Searle factors the "slogan ... 'the mind is to the brain as the program is to the hardware' " (p. 21) into three questions:

1. Is the brain a digital computer?
2. Is the mind a computer program?
3. Can the operations of the brain be simulated on a digital computer? (Searle 1990: 21.)

Let us consider each of these, beginning with the second.

8.3.1 Is the Mind a Computer Program?

What does it mean to say that the mind is a computer program? Surely not that there is a programming language and a program written in it that is being executed on a brain—not for humans, at least. So it could mean that by bottom-up, reverse engineering (neuroscience) together with top-down, cognitive-scientific investigation, we could write a program that would cause a computer to exhibit mental behavior. But that's question 3, to which Searle gives a different answer.

Possibly question 2 means that the mind plays the same role with respect to the brain that a program does to a computer, what I've called Good Old-Fashioned Cartesian Dualism (§7.2). But that's not much progress over the "slogan" of which question 2 is supposed to be merely a part.

[1] I owe some of these points to David A. Zubin (in conversation) and to Johan Lammens (personal communication).

Does question 2 mean that the mind is the *way* the brain behaves? That seems right, but isn't the right parallel: It *doesn't* seem right to say that a program is the way a computer behaves.

"Programs," Searle goes on to say, "are defined purely formally or syntactically" (p. 21). That, I think, is not *quite* right: They require a set of input–output conventions, which would be "links" to the world. In any case, this together with the assertion that "minds have an intrinsic content ... immediately [implies] that the program by itself cannot constitute the mind" (p. 21). What does 'content' mean? If it means something internal to the mind (a "container" metaphor; cf. Twardowski 1894, Rapaport 1978), then that minds have intrinsic content could mean that within a mind there are links among nodes, some of which play the role of a language of thought and others of which play the role of mental representations of external perceptions (§3.2.2.2.1). If so, that would be—as Searle says programs are—purely syntactic.

If, on the other hand, 'content' means a relation to an external entity, then why don't programs have that, too (as we just noted)? In any case, programs do take input from the external world: I enter '2' on the keyboard, which results (after a few transductions) in a switch being set in the computer, which the program interprets as the number 2.

So, on either interpretation, the conclusion doesn't follow, since programs can *also* have "intrinsic mental content", whatever that means.

The problem is that question 2 is not the right question. Of course "The formal syntax of the program does not by itself guarantee the presence of mental contents" (p. 26), because the program might never be executed. What Searle should have asked is whether the mind is a computer *process*. And here the answer can be 'yes', since *processes* can have contents.

> I showed this [viz., that the formal syntax of a program doesn't guarantee the presence of mental contents] a decade ago in the CRA **???** .... The argument rests on the simple logical truth that syntax is not the same as, nor is it by itself sufficient for, semantics. (Searle 1990: 21.)

Well, by Morris's definitions (1938), syntax $\neq$ semantics. Sure. Nor is it the case that semantics can be "derived", "constructed", or "produced" from syntax by Morris's definitions. But the first-person semantic enterprise *is* one of determining correspondences among symbols—between linguistic symbols and internal representations of external objects. Hence, it *is* syntactic even on Morris's definition. The *third*-person semantic enterprise is more like what Morris had in mind. But one person's third-person semantics is another's first-person semantics: If Oscar tries to account for Cassie's semantics by drawing correspondences between her nodes and things in the world, all he can really do is draw correspondences between his representations of her nodes and his representations of things in the world. As with the turtles, it's syntax all the way down.

8.3.2 Can the Operations of the Brain Be Simulated on a Digital Computer?

Let's turn to question 3, the answer to which Searle thinks is trivially—or, at least, uninterestingly—affirmative. "[N]aturally interpreted, the question means: Is there some description of the brain such that under that description you could do a computational simulation of the operations of the brain" (p. 21). Such a description would be like one of Smith's models. Following Smith, then, we would have to claim that such a model would be partial. Hence, so would be the computational simulation. But if it passed the Turing test (i.e., if its effects in the actual world were indistinguishable from those of a human), then what's not in the model is an implementation detail. What might these be? Consistent with our results in §7.6.3 on qualia, they might include sensations of pain, warm fuzzy feelings associated with categorizing something as "beautiful", etc. As for pain, don't forget that our *sensation* of it is an internal perception, just like our sensation of an odor. It might be possible to be in pain and to know that one is in pain without what we normally call a pain sensation, just as it is possible to determine the presence of an object by its odor—by a chemical analysis—without sensing that odor.[2] The "triviality" or "obviousness" of the answer to question 3 stems, according to Searle, from Church's Thesis: "The operations of the brain can be simulated on a digital computer in the same sense in which weather systems, the behavior of the New York Stock market or the pattern of airline flights over Latin America can" (p. 21). And, presumably, since simulated weather isn't weather, simulated brains aren't brains. But the premise is arguable (§7.6.4); at least, it does not follow that the behavior of simulated brains isn't *mental*. Brains and brain behavior are special cases.

8.3.3 Is the Brain a Digital Computer?

Searle equates question 1 to "Are brain processes computational?" (p. 22). What would it mean to say that the brain was *not* a digital computer? It might mean that the brain is *more* than a digital computer—that only some proper *part* of it *is* a digital computer. What would the rest of it be? Implementation details, perhaps. I am, however, willing to admit that perhaps not all of the brain's processes are computational. Following Johnson-Laird (1988: 26–27), I take the task of cognitive science to be to find out *how much* of the brain's processes *are* computational—and surely some of them are. It is, thus, a working hypothesis that brain processes are computational, requiring an empirical answer and not subject to apriori refutation.

On the other hand, to say that the brain is not a digital computer might mean that it's a different kind of entity altogether—that no part of it is a digital computer. But that seems wrong, since it *can* execute programs (we use our brains to hand-simulate computer programs).

What are brain processes, how do they differ from mental processes, and how do both of these relate to computer processes? A computer process is a program being executed; therefore, it is a physical thing that implements an abstract program. A brain process is also a physical thing, so it would seem to correspond to a computer process. A mental process could be either (i) something abstract yet dynamic or (ii) a brain process. The former (i) makes no sense if programs and minds are viewed as static entities. The latter (ii) would mean that *some* brain processes are mental (others, like raising one's arm, are not). So to ask if brain processes are computational is like asking if a computer process is computational. That question means: Is the current behavior of the computer describable by a recursive function (or is it just a fuse blowing)? So Searle's question 1 is: Is the current (mental) behavior of the brain describable by a recursive function? This is the fundamental question of artificial intelligence as computational philosophy. It is a major research program, not a logical puzzle capable of apriori resolution.

Searle's categorization of the possible positions into "strong AI" ("all there is to having a mind is having a program") "weak AI" ("brain processes (and mental processes) can be simulated computationally"), and "Cognitivism" ("the brain is a digital computer") is too coarse (p. 22). **[??? check quotations]** What about the claim that a computer running the "final" AI program (the

[2] Angier 1992: A19 reports that "Sperm cells possess the same sort of odor receptors that allow the nose to smell." This does not mean, of course, that sperm cells have the mental capacity to have smell-qualia. And Blakeslee 1993 reports that "humans ... may exhude ... odorless chemicals called pheromones that send meangful signals to other humans." She calls this "a cryptic sensory sytstem that exists without conscious awareness ...."

one that passes the Turing test, let's say) has mentality? As I argued above, that's not necessarily "just" having a program. But if the *process* interpretation of question 2 is taken, then Strong AI could be the view that all there is to having a mind is having a *process*, and that's more than having a program. What about the claim that the "final" AI program need not be the one that humans use—i.e., the claim that computational *philosophy* might "succeed", not computational *psychology* (cf. Rapaport 1986a, Shapiro 1992a)? This is a distinction that Searle does not seem to make. Finally, Pylyshyn's version of "cognitivism" (1985) does not, I think, claim that the brain *is* a digital computer, but that mental processes are computational processes. That seems to me to be compatible with the brain being "more" than a digital computer.

As we saw (§7.7.1), Searle complains that multiple realizability is "disastrous" (p. 26). The first reason is that *anything* can be described in terms of 0s and 1s (p. 26). And there might be *lots* of 0–1 encodings of the brain. But the real question, it seems to me, is this: Does the brain *compute* (effectively) some function? What is the input–output description of that function? The answer to the latter question is whatever psychology tells us is intelligence, cognition, etc. For special cases, it's easier to be a bit more specific: For natural-language understanding, the input is some utterance of natural language, and the output is an "appropriate" response (where the measure of "appropriateness" is defined, let's say, sociolinguistically). For vision, the input is some physical object, and the output is, again, some "appropriate" response (say, an utterance identifying the object or some scene, or some behavior to pick up or avoid the object, etc.). Moreover, these two modules (natural-language understanding and vision) must be able to "communicate" with each other. (They might or might not be modular in Fodor's sense (1983), or cognitively impenetrable in Pylyshyn's sense (1985). In any case, solving one of these problems will require a solution to the other; they are "AI-complete" (Shapiro 1992a).)

The second allegedly disastrous consequence of multiple realizability is that "syntax is not intrinsic to physics. The ascription of syntactical properties is always relative to an agent or observer who treats certain physical phenomena as syntactical" (p. 26). The observer assigns 0s and 1s to the physical phenomena. But Morris's definition of syntax as relations among symbols (uninterpreted marks) can be extended to relations among components of any system. Surely, physical objects stand in those relationships "intrinsically". And if 0s and 1s *can* be ascribed to a physical object (by an observer), that *fact* exists independently of the agent who *discovers* it.

Searle's claim "that syntax is essentially an observer relative notion" (p. 27) is very odd. One would have expected him to say that about *semantics*, not syntax. Insofar as one can look at a complex system and describe (or discover) relations among its parts (independently of any claims about what it does at any higher level), one is doing *non*–observer-relative syntax. Searle says that "this move is no help. A physical state of a system is a computational state only relative to the assignment to that state of some computational role, function, or interpretation" (p. 27), where, presumably, the assignment is made by an observer. But an assignment is an assignment of meaning; it's an interpretation. So is Searle saying that computation is fundamentally a *semantic* notion? But, for Church, Turing, et al., computation is purely *syntactic*. It's only the input–output coding that *might* constitute an assignment. But such coding is only needed in order to be able to link the syntax with the standard theory of computation in terms of functions from natural numbers to natural numbers. If we're willing to express the theory of computation in terms of functions from physical states to physical states (and why shouldn't we?), then it's not relative.

Searle rejects question 1: "There is no way you could discover that something is intrinsically a digital computer because the characterization of it as a digital computer is always relative to an observer who assigns a *syntactical* interpretation to the purely physical features of the system" (p. 28, my italics). I, too, reject question 1, but for a very different reason: I think the question is really whether *mental processes* are computational. In any event, suppose we *do* find computer programs that exhibit intelligent input–output behavior, i.e., that pass the Turing Test. Computational *philosophy* makes no claim about whether that tells us that the human *brain* is a digital computer. It only tells us that intelligence is a computable function. So at best Searle's arguments are against computational *psychology*. But even that need not imply that the brain *is* a digital computer, only that it behaves as if it were. To discover that something X is intrinsically a digital computer, or a Y, is to have an abstraction Y, and to find correspondences between X and Y.

Perhaps what Searle is saying is that being computational is not a *natural* kind, but an artifactual kind (cf. Churchland & Sejnowski 1992):

> I am not saying there are *a priori* limits on the patterns we could discover in nature. We could no doubt discover a pattern of events in my brain that was isomorphic to the implementation of the vi program on this computer. (Searle 1990: 28.)

This is to admit what I observed two paragraphs back. Searle continues:

> But to say that something is *functioning as* a computational process is to say something more than that a pattern of physical events is occurring. It requires the assignment of a computational interpretation by some agent. (Searle 1990: 28.)

But why? Possibly because to find correspondences between two things (say, a brain and the Abstraction ComputationalProcess—better, the Abstraction Computer) is observer-relative? But if we have *already* established that a certain brain process is an implementation of `vi`, what *extra* "assignment of a computational interpretation by some agent" is needed?

Searle persists:

> Analogously, we might discover in nature objects which had the same sort of shape as chairs and which could therefore be used as chairs; but we could not discover objects in nature which were functioning as chairs, except relative to some agent who regarded them or used them as chairs. (Searle 1990: 28.)

The analogy is clearly with *artifacts*. But the notion of a computational process does not seem to me to be artifactual; it is *mathematical*. So the proper analogy would be something like this: Can we discover in nature objects that were, say, sets, or numbers, or Abelian groups? Here, the answer is, I think, (a qualified) 'yes'. (It is qualified, because sets and numbers are abstract and infinite, while the world is concrete and finite. Groups may be a clearer case.) In any event, is Searle claiming that the implementation of `vi` in my brain isn't `vi` until someone *uses it as* `vi`? If there is an implementation of `vi` on my Macintosh that no one ever uses, it's still `vi`.

Searle accuses computational cognitive scientists of "commit[ing] the homunculus fallacy ... treat[ing] the brain as if there were some agent inside it using it to compute with" (p. 28). But recall Hayes's objection to the Chinese-Room Argument: Computation is a series of switch-settings; it isn't rule-following. (On this view, by the way, the solar system *does* compute certain mathematical functions; see below). Turing machines do *not* follow rules; they simply change state. There are, however, descriptions—programs—of the state changes, and anything that follows (executes) that program computes the same function computed by the Turing machine. A *universal* Turing machine can also follow that program. But the original, special-purpose Turing machine's program is "hardwired" (an analogy, of course, since everything is abstract here). A universal Turing machine has *its* program similarly hardwired. It is only when the universal Turing machine is fed a program that it follows the rules of that program. But that's what *we* do when *we* consciously follow (hand-simulate) the rules of a program. So it's *Searle* who commits the homuncular fallacy in the Chinese-Room Argument by putting a person in the room. It is not the person in the room who either does or does not understand Chinese; it is the entire system. Similarly, it is not some *part* of my brain that understands language; it is *I* who understands.

In his discussion of "discharging" the homunculus, Searle says that "All of the higher levels reduce to this bottom level. Only the bottom level really exists; the top levels are all just *as-if*" (p. 29). But as I have argued elsewhere (Rapaport 1990), *all* levels exist, and *all* levels "do the same thing" (albeit in different ways).

I noted above that systems that don't follow rules can still be said to be computing. My example was the solar system. Searle offers "nails [that] compute the distance they are to travel in the board from the impact of the hammer and the density of the wood" (p. 29) and the human visual system; "neither," according to him, "compute anything" (p. 29). But in fact they both do. (The nail example might not be ideal, but it's a nice example of an *analog* computation.)

> But you do not *understand* hammering by supposing that nails are somehow intrinsically implementing hammering algorithms and you do not *understand* vision by supposing the system is implementing, e.g., the shape from shading algorithm. (Searle 1990: 29; my italics.)

Why not? It gives us a theory about how the system might be performing the task. We can falsify (or test) the theory. What more could *any* (scientific) theory give us? What further kind of understanding could there be? Well, there could be first-person understanding, but I doubt that we could ever know what it is like to be a nail or a solar system. We *do* understand what it is like to be a cognitive agent!

The problem, I think, is that Searle and I are interested in different (but complementary) things:

> ... you cannot explain a physical system such as a typewriter or a brain by identifying a pattern which it shares with its computational simulation, because the existence of the pattern does not explain how the system actually works *as a physical system*. (Searle 1990: 32.)

Of course not. That would be to confuse the implementation with the Abstraction. Searle is interested in the former; he wants to know how the *brain* works. I, however, want to know what the brain *does* and how *anything* could do it. For that, I need an account at the functional/computational level, not a biological (or neuroscientific) theory.

> The mistake is to suppose that in the sense in which computers are used to process

> information, brains also process information. [Cf. Johnson 1990.] To see that that is a mistake, contrast what goes on in the computer with what goes on in the brain. In the case of the computer, an outside agent encodes some information in a form that can be processed by the circuitry of the computer. That is, he or she provides a syntactical realization of the information that the computer can implement in, for example, different voltage levels. The computer then goes through a series of electrical stages that the outside agent can interpret both syntactically and semantically even though, of course, the hardware has no intrinsic syntax or semantics: It is all in the eye of the beholder. And the physics does not matter provided only that you can get it to implement the algorithm. Finally, an output is produced in the form of physical phenomena which an observer can interpret as symbols with a syntax and a semantics.
>
> But now contrast this with the brain. ... none of the relevant neurobiological processes are observer relative ... and the specificity of the neurophysiology matters desperately. (Searle 1990: 34. **[??? check paragraphing]**)

There is much to disagree with here. First, "an outside agent" need *not* "encode ... information in a form that can be processed by the circuitry of the computer". A computer could be (and typically is) designed to take input directly from the real world and to perform the encoding (better: the transduction) itself, as, e.g., in document-image understanding (cf. Srihari & Rapaport 1989, 1990; Srihari 1991ab, 1993ab). Conversely, abstract concepts are "encoded" in natural language so as to be processable by *human* "circuitry".

Second, although I of course find the phrase 'syntactical realization' quite congenial (cf. Ch. 2), I'm not sure how to parse the rest of the sentence in which it appears. What does the computer implement in voltage levels: the information? The syntactical realization? I'd say the former, and that the syntactical realization *is* the voltage levels. So there's an issue here of whether the voltage levels are *interpreted as* information, or vice versa.

Third, the output need not be physical phenomena interpreted by an observer as symbols. The output *could* be an action, or more internal data (e.g., as in a vision system),[3] or even natural language to be interpreted by another *computer*. Indeed, the latter suggests an interesting research project: Set up Cassie and Oscar, our computational cognitive agents implemented in SNePS. Let Cassie have a story pre-stored or as the result of "reading" or "conversing". Then let her tell the story to Oscar and ask him questions about it. No *humans* need be involved.

Fourth, neurobiological processes aren't observer-relative only because we don't care to, or need to, describe them that way. The computer works as it does independently of us, too. Of course, for *us* to understand what the brain is doing—from a third-person point of view—we need a psychological level of description (cf. Chomsky 1968, Fodor 1968).

Finally, why should "the specificity of the neurophysiology matter desperately"? Does this mean that if the neurophysiology were different, it wouldn't be a human brain? I suppose so, but that's relevant only for the implementation side of the issue, not the Abstraction side, with which I am concerned.

Here is another example of how Searle does not seem to understand what computational cognitive science is after:

[3]Searle seems to think (p. 34) that vision systems yield sentences as output! (See below.)

> A standard computational model of vision will take in information about the visual array on my retina and eventually print out the sentence, "There is a car coming toward me". But that is not what happens in the actual biology. In the biology a concrete and specific series of electro-chemical reactions are set up by the assault of the photons on the photo receptor cells of my retina, and this entire process eventually results in a concrete visual experience. The biological reality is not that of a bunch of words or symbols being produced by the visual system, rather it is a matter of a concrete specific conscious visual event; this very visual experience. (Searle 1990: 34–35.)

The first sentence is astounding. First, why does he assume that the input to the computational vision system is *information on the retina*, rather than *things in the world*? The former is close to an *internal* symbol representing external information! Second, it is hardly "standard" to have a vision system yield a *sentence* as an output. It might, of course ("Oh, what a pretty blue flower."), but, in the case of a car coming at the system, an aversive maneuver would seem to be called for, not a matter-of-fact description. Nonetheless, precisely that input–output interaction *could*, *pace* Searle, be "what happens in the actual biology": I could say that sentence upon appropriate retinal stimulation.

Of course, as the rest of the quotation makes clear, Searle is more concerned with the intervening qualitative experience, which, he seems to think, humans have but computers don't (or can't). Well, could they? Surely, there ought to be an intervening stage in which the retinal image is processed (perhaps stored) before the information thus processed or stored is passed to the natural-language module and interpreted and generated. Does that process have a qualitative feel? Who knows? *How* would you know? Indeed, how do I know (or believe) that *you* have such a qualitative feel? The question is the same for both human and computer. As with Shapiro's pain-feeling computer (§7.6.3), it's possible that a physical theory of sensation could be constructed. Is it computational? Perhaps not—but so what? As I urged in §7.6.3, perhaps some "mental" phenomena are not *really* mental (or computational) after all. Or perhaps a computational theory will always be such that there is a role to play for some sensation or other, even though the actual sensation in the event is not computational. That is, every computational theory of pain or vision or what have you will be such that it will refer to a sensation without specifying what the sensation is. (Cf. Shoemaker on qualia. **[???]** Cf., also, Gracia's example of a non-written universal for a written text, §**??**.) **[???]**

Of course, despite my comments about the linguistic output of a vision system, the sentence that Searle talks about could be a "sentence" of one's language of thought. That, however, would fall under the category of being a "concrete specific conscious visual event" and "not ... a bunch of words or symbols" (cf. Pylyshyn 1981; Srihari & Rapaport 1989, 1990; Srihari 1991ab, 1993ab).

Searle's final point about question 1 is this:

> The point is not that the claim "The brain is a digital computer" is false. Rather it does not get up to the level of falsehood. It does not have a clear sense. (Searle 1990: 35.)

This is because "you could not *discover* that the brain *or anything else* was intrinsically a digital computer" (p. 35, my italics). "Or anything else"? Even an IBM PC? Surely not. Possibly he means something like this: Suppose we find an alien physical object and theorize that it is a digital computer. Have we *discovered* that it is? No—we've got an *interpretation* of it *as* a digital computer (cf. "you could assign a computational interpretation to it as you could to anything else" (p. 35)). But how else *could* we "discover" anything about it? Surely, we could discover that it's made of silicon and has 10^k parts. But that's consistent with his views about *artifacts*. Could we *discover* the topological arrangement of its parts? I'd say 'yes'. Can we *discover* the sequential arrangement of its behaviors? Again, I'd say 'yes'. Now consider this: How do we determine that it's made of silicon? By subjecting it to certain physical or chemical tests and having a theory that says that any substance that behaves thus and so is (made of) silicon. But if anything that *behaves* such and thus is a computer, then so is this machine! So we *can* discover that (or whether) it is a computer. (Better: We can discover whether its processing is computational.)

8.4 RETURN TO THE HELEN KELLER ROOM.

> As we returned to the house every object which I touched seemed to quiver with life. That was because I saw everything with the strange, new sight that had come to me. (Keller 1905: 36.)

Thus Helen Keller described her experience immediately after the well-house episode. And this, I claim, would eventually be the experience of Searle-in-the-room, who would then have *semantic* methods for doing things in addition to purely *syntactic* ones (just as logicians have both syntactic and semantic methods of proof). The semantic methods, however, are strictly internal—correspondences among internal nodes for words and things.

Jim Swan (1991) points out how important Helen's *hand* is to her ability to communicate. In the well-house episode, both the name and the object were communicated via the *same* sense modality: touching her hand. [**POSSIBLY CITE P.2 OF SWAN 1991**]. He also points out how she had to learn about the visual dimension of the world *as language*. All of this lends credence to the view that Helen's understanding of language and the world was an internal understanding. Nodes for both words and things were built on the basis of tactile (and olfactory) sensation. One of the reasons the well-house episode was significant was that it was the event that enabled Helen to distinguish some of her internal nodes from others, categorizing some as representing the world and others as names of the former. For Helen, initially, language was indistinguishable from the non-linguistic part of the world.

Swan discusses, from a psychoanalytical point of view, the difficulty for Helen of distinguishing between self and other, between her words and those of others.[4] Before the well-house episode, she could use signs, but had difficulties with some, in particular, with those for container vs. contained ('milk' or 'water' vs. 'mug'; see Ch. 9). Perhaps, before, she could not distinguish words from things: Words *were* things, part of a holistic fabric of the world. Afterwards, she could distinguish between two kinds of things in the world: things and words for them. That placed a syntactic–semantic structure on her mental network. And it resulted, as we know, in the blossoming of her understanding. Searle-in-the-room could do no worse.

Before the well house, Helen used symbols to communicate, but not to think. Harman, recall, said that

[4] Helen had been accused of plagiarism, when, in fact, it is possible that she had merely made a grievous use–mention confusion, viz., not having learned how to use quotation marks; cf. [ref] **either Keller 1905 passim or Swan**.

> a language, properly so called, is a symbol system that is used both for communication and thought. If one cannot think in a language, one has not yet mastered it. A symbol system used only for communication, like Morse code [or, one might add, like Helen's pre–well-house signs], is not a language. (Harman 19xx: 00.**???**)

Unless the symbols are part of a larger network, they have no (or very little) meaning. To that extent, perhaps Searle has a point. But the more they *are* used for thinking, the more language-like they are. And they *have* to be part of a larger network—partitioned into syntactic and semantic regions—else they could not be used to communicate. They have meaning if and only if (and to the extent that) they are part of such a larger, partitioned network. Searle denies the "if" part of this, but Helen Keller, I suggest, was a living counterexample.

Chapter 9

NAMES FOR THINGS: FROM "MONKEY-LIKE IMITATION" TO NATURAL-LANGUAGE UNDERSTANDING.

9.1 A PUZZLE.

I have suggested that the case of Helen Keller offers a real-life Chinese Room situation (§8.2.3), and I have given some reasons why the epiphenal well-house episode—paradigmatic of the syntax–semantics relationship, with Helen simultaneously having one hand immersed in syntax and the other in semantics—was so significant for her (§8.4).

But, really, why should it have been? By Helen's and Annie Sullivan's own testimony, Helen seemed able to use (finger-spelled) words for things, as well as (self-invented) signs and gestures for things, *before* the well house. So what made the well house so significant?

9.2 WHAT DID HELEN KELLER UNDERSTAND, AND WHEN DID SHE UNDERSTAND IT?

It is not easy to determine the chronology of Helen's language learning. There are two distinct, if not independent, first-person accounts: (1) the student's: Helen's autobiography (Keller 1905), written, of course, long after the events,[1] and (2) the teacher's: Annie Sullivan's contemporaneous letters.[2] The latter are probably to be trusted more. As Helen herself says, "When I try to classify my earliest impressions, I find that fact and fancy look alike across the years that link the past

[1] There are also Helen's letters, but these, while intrinsically interesting and exhibiting—especially in the early ones—her gradual mastery of language, do not contain much information on *how* she learned language.

[2] There are also Annie's speeches and reports. Although they contain some useful information and some valuable insights—especially into the nature of teaching—they, like Helen's autobiography, are *ex post facto*; cf. Macy, in Keller 1905: 278.

Figure 9.1:

with the present. The woman paints the child's experiences in her own fantasy" (Keller 1905: 23; cf. p. 224).[3] Even though Annie Sullivan's letters are "incomplete" as scientific "records" (Macy, in Keller 1905: 239; cf. p. 241), they are the closest thing available. Together, the two sources provide a reasonable—if tantalizing—picture.

For Helen's earliest years, we must rely on her own report. She was born, healthy, on 27 June 1880. At 6 months, she could speak a few words or phrases, e.g., 'How d'ye', 'tea', and—significantly—'wa-wa' ("water") (p. 25). (But did she understand them? See Figure 9.1.) Sometime after her first birthday **[FIND OUT WHEN; SEE THE *BIOGRAPHY* OF HER]**, she contracted an illness that left her deaf and blind, and, like many deaf children, she did not learn to speak. Nevertheless, she could make sounds—again, significantly, the sound 'wa-wa' for "water", which "I ceased making … only when I learned to spell the word" (p. 25). After her recovery, she could communicate via touch, "crude signs" (shaking her head for 'no', nodding for 'yes', pulling for 'come', pushing for 'go'; cf. p. 27), and (other) imitative motions, including some rather complex ones (p. 28).

[3]This has an overtone of holistic reinterpretation; cf. §2.8.2: We understand the present in terms of all that has gone before, and the past in terms of all that has come after.

She familiarized herself with the outdoors, guided by her sense of smell. It is perhaps worth noting that this continued well after Annie Sullivan's arrival: They often studied outside (p. 43). (Cf., too, pp. 293ff on the significance of the sense of smell.) She also, of course, had a sense of taste, learning thereby that the ocean was salty—a bit of commonsense knowledge that she lacked because no one thought to tell her such an obvious thing (p. 230)! She also had a "sense" of vibration, being able to sense when a door closed (p. 27). I am not sure whether to count this as part of her sense of touch, or as a remnant of a sense of hearing, which is, after all, a sensitivity to vibrations (cf. p. 208 on her ability to sense music).

By the age of 5, she could perform rather complex tasks, such as folding and separating clean clothes, and she knew that her father did something mysterious by holding a newspaper in front of his eyes. She imitated this, but it did not illuminate the mystery (p. 30; cf. the "miracle of reading", §5.1). Similarly, she knew that others communicated, not with signs, but by moving their lips; imitation of this, too, was not successful (pp. 27–28).

Fortunately, others understood her signs (p. 28). When she was 6, she may have tried to teach her dog some of these (with no success, of course; p. 29), though Annie Sullivan tells a similar story about Helen trying to teach *finger spelling* to her dog at about the same age or a bit later (20 March 1887, to be exact—*after* Helen had begun to learn words but *before* the well house). She certainly, at about this time, had a desire to express herself (p. 32).

On 3 March 1887, Annie Sullivan arrived at Helen's home to become her teacher; Helen was now 3 months shy of 7 years old. Almost immediately upon her arrival, Annie Sullivan and Helen began to communicate with each other using signs and gestures (p. 24). The next day, Annie Sullivan began teaching Helen finger spelling, presenting her with an object or action and finger-spelling its name: 'doll', 'pin', 'hat', 'cup', 'sit', 'stand', and 'walk' are the words Helen remembered. Annie Sullivan cites 'doll', 'cake', and (sewing) 'card'. 'Cup' is of some interest, since 'mug' was to give Helen a notorious difficulty a few weeks later.

To what extent did Helen understand these words? As we saw, she herself considered this to have been "monkey-like imitation" (p. 35): Finger spelling was an activity to be performed upon presentation of certain objects. It was a ritual, with a syntactic structure: There were right and wrong ways to perform it. But Helen did this "in ... [an] uncomprehending way" (p. 35) and did not yet understand "that everything has a name" (p. 35).

Was Helen really so uncomprehending at this stage? Recall that she had already developed her own system of signs and gestures for communicating her needs and wants. Surely, this is evidence of a semantic correspondence. I suppose it is remotely possible that even Helen's early self-invented signs were ritual movements performed uncomprehendingly in certain circumstances, yet rituals that just happened to convey information to other people. (In Robert Sheckley's story "Ritual" (1954), creatures living on a remote planet perform a series of ritual "welcoming the gods" dances as a religious ceremony. The dance happens to consist of the preparations for the arrival of a spaceship. When a spaceship finally does arrive after centuries without a landing, the villagers perform their "dance", which just happens to facilitate the spaceship landing.) But I doubt that Helen's signs were such rituals. Had *all* of Helen's gestures been such conveniently coincidental ("extensional") rituals, she would not have been able to do the complex tasks she did, or to satisfy her needs, or to have the appropriate background knowledge that, eventually, was the basis for her language learning.

All that Annie Sullivan was doing can be seen as offering Helen a new system for accomplishing her communicational goals. It is, of course, possible that Helen did not realize this, so that, for her, her *own* gestures for an object *did* constitute a semantic correspondence while Annie Sullivan's finger spellings did not. However, that Helen *was* able to associate finger spellings with objects and actions surely indicates that she had the means to promote these to *semantic* correspondences.

There is, in fact, evidence that she did so: The day that Annie Sullivan arrived, she taught Helen 'cake', and the next day she taught her 'card'. Helen ...

> ... made the "c-a," then stopped and thought, and making the sign for eating and pointing downward she pushed me [Annie Sullivan] toward the door, meaning that I must go downstairs for some cake. The two letters "c-a," you see, had reminded her of Friday's "lesson"—not that she had any idea that *cake* was the name of the thing, but it was simply a matter of association, I suppose. (Keller 1905: 246.)

I would argue that Helen *did* have the idea that 'cake' "was the name of the thing"—but that she had that idea *de re*, not *de dicto*: She did not yet have the *concept* of names for things. She could certainly associate words (i.e., finger spellings) with objects:

> Then I [Annie Sullivan] spelled "d-o-l-l" and began to hunt for it. She [Helen] follows with her hands every motion you make, and she knew that I was looking for the doll. She pointed down, meaning that the doll was downstairs. ... [S]he ran downstairs and brought the doll ... (Keller 1905: 246–247.)

although not without a reward of some cake.

As Helen built up a vocabulary of finger-spelled words and made mental links of (her internal representations of) these with (her internal representations of) things and actions, she was building a "semantic network" of associated representations that she could and did use in a linguistic (or language-like) way. Searle would argue that she did not *understand* language. Perhaps. I'd prefer to say that she did not understand language in a *de dicto* way—she *did* understand it *de re*, in the sense that she was using it, but did not realize that she was using it or how it worked. She was, thus, at the same stage of language development as a normal child would have been at a much earlier age. Are we prepared to say that normal children at this stage do not understand language? Perhaps. But eventually they *do*, and eventually Helen did. Why not Searle-in-the-room or a computer? What is the crucial step (or steps) that must be taken to move from *this* level of understanding (or, if you prefer, from this level of *not* understanding) to the level that we adult speakers of language are at? We'll return to this in §9.4.

By 11 March 1887, Annie Sullivan says that "Helen knows several words now, but has no idea how to use them, or that everything has a name" (p. 251). Yet two days later, Helen can associate words with objects: "when I give her the objects, the names of which she has learned, she spells them unhesitatingly" (p. 251).

On 20 March 1887, Helen reports that she was confused by 'mug' and 'water', apparently not being able to distinguish the container from the contained (p. 36). Perhaps this was because they always appeared together. Although she may have had a *mug* by itself, perhaps the *water* was always *in* the mug. Here are Annie Sullivan's accounts:

> Helen has learned several nouns this week. "M-u-g" and "m-i-l-k" have given her more trouble than other words. When she spells "milk," she points to the mug, and when she spells "mug," she makes the sign for pouring or drinking, which shows that she has confused the words. She has no idea yet that everything has a name. (Annie Sullivan, 20 March 1887, in Keller 1905: 252–253.)

By 'learning' (as in "Helen has learned several nouns this week"), Annie Sullivan must mean the ability to spell, to make the finger movements. 'Mug' and 'milk' (or 'water'?) give Helen trouble, *not* in terms of the movements, but in terms of how to use them (what they refer to, or name). But that assumes that Helen knows that they have a *use*, which, although not at all clear, is at least plausible, as we've seen. Of course, pointing to the mug *could* also be pointing to the milk (or water) *in* the mug. Helen's own version (p. 36) suggests that she was not making *any* distinctions at all, rather than merely confusing the mug and the liquid. Annie Sullivan's second version of the confusion supports my interpretation that Helen was aware only of *events* considered as unanalyzed wholes:

> ... "mug" and "milk" had given Helen more trouble than all the rest. She confused the nouns with the verb "drink." She didn't know the word for "drink," but went through the pantomime for drinking whenever she spelled "mug" or "milk". (Annie Sullivan, 5 April 1887, in Keller 1905: 256.)

I think it is significant that Annie Sullivan reported the confusion as between 'mug' and '*milk*', where Helen reported it as between 'mug' and '*water*'.[4] First, and most importantly (if only for Freudian reasons), Helen's one remaining *spoken* word was, you will recall, 'water' ('wa-wa'). Second, if Annie Sullivan's report is the one to be trusted, besides the semantic-domain confusion between container and contained, there might also have been a syntactic-domain confusion between two words beginning with 'm': Recall the earlier "confusion" between 'ca[ke]' and 'ca[rd]'.

There were a few days to go before the visit to the well house. What did Helen learn in those days between her confusing the words for a mug and its liquid contents and her later epipheny? By the 20th of March, according to Annie Sullivan, Helen knew 12 word–object combinations (p. 255), yet instinctively used her own signs—not finger-spelled words—to *communicate*. By 1 April, Annie Sullivan reports that Helen's vocabulary had increased to 25 nouns and 4 verbs[5]—including, significantly, 'mug', 'milk', and 'water'. Yet, two days later, Annie Sullivan says, Helen "has no idea what the spelling means" (p. 256). I take it that, from Annie Sullivan's point of view, Helen's "knowledge" of these words was at least associative and probably even communicative, yet not "conscious". But not "conscious" in what sense? Helen apparently could ask for the finger spellings that corresponded to certain objects (the ones marked 'x' in note 5.) What more could Annie Sullivan want at this stage?

Searle, no doubt, would say that for real natural-language understanding, a lot more is wanted. I'd have to agree: Helen could not yet have passed a Turing test. So although imagining what Helen was like at this stage may give us an insight as to what Searle-in-the-room is like, there is a large gap between the two. Searle-in-the-room, remember, passes the Turing test.

[4]In an earlier autobiography, Helen also called this a 'mug'/'*milk*' confusion (Keller 1905: 364).

[5]"*Doll, mug, pin, key, dog, hat, cup, box, water, milk, candy, eye (x), finger (x), toe (x), head (x), cake, baby, mother, sit, stand, walk. ... knife, fork, spoon, saucer, tea, paper, bed,* and ... *run*" (p. 256). ("Those with a cross after them are words she asked for herself" (p. 256).)

Perhaps what Helen "knew" at this stage was an association of these words with certain complex, unanalyzed events, and that what she learned at the well house was that the events have parts, each of which is associated with a word. If so, then what she learned was as much about the semantic domain as it was about the association between the two domains. Of course, she also presumably learned then that the *words* did not refer to complex events but only to *parts* of them. So she learned something about the syntactic domain, too.

9.3 THE WELL HOUSE: EPIPHENY, PUZZLE, AND SIGNIFICANCE.

9.3.1 Epiphany.

The magical day was 5 April 1887. Annie Sullivan, having failed to clarify the difference between 'mug' and 'milk', took Helen for a walk to the well house.

> This morning, while she was washing, she wanted to know the name for "water." ... I spelled "w-a-t-e-r" .... [I]t occurred to me that with the help of this new word I might succeed in straightening out the "mug–milk" difficulty. We went out to the pump-house, and I made Helen hold her mug under the spout while I pumped. As the cold water gushed forth, filling the mug, I spelled "w-a-t-e-r" in Helen's free hand. The word coming so close upon the sensation of cold water rushing over her [other] hand seemed to startle her. She dropped the mug and stood as one transfixed. A new light came into her face. She spelled "water" several times. Then she dropped on the ground and asked for its name and pointed to the pump and the trellis, and suddenly turning round she asked for my name. I spelled "Teacher." Just then the nurse brought Helen's little sister into the pump-house, and Helen spelled "baby" and pointed to the nurse. All the way back to the house she was highly excited, and learned the name of every object she touched, so that in a few hours she had added thirty new words to her vocabulary. ...
>
> ... Helen got up this morning like a radiant fairy. She has flitted from object to object, asking the name of everything and kissing me for very gladness. Last night when I got in bed, she stole into my arms of her own accord and kissed me for the first time, and I thought my heart would burst, so full was it of joy. (Annie Sullivan, 5 April 1887, in Keller 1905: 256–257.)

A few observations on this passage and on the well-house episode are in order. First, clearly, Helen wanted to know the name for water, not for 'water'; she did not want to know the name for a name. However, Annie Sullivan is not to be blamed for this particular use–mention confusion! On the other hand, hasn't Annie Sullivan repeatedly told us that Helen did not know that things have names? Then why does she report Helen as asking for the *name* of water? Perhaps this needs to be taken *de re*. Note that it's possible that what Helen wanted to know was the appropriate finger spelling for *washing*.

Second, Annie's comment about "straightening out the 'mug–milk' difficulty" can be intepreted as supporting my suggestion that the mug–milk confusion was one of a container vs.

contained or of unanalyzed events.

Third, note that here there was little chance to "confound" two objects—there was a direct and simultaneous association of word with object. Although the mug in Helen's hand *might* have caused some interference, Helen's own account indicates that it did not:

> We walked down the path to the well-house .... Some one was drawing water and my teacher placed my hand under the spout. As the cool stream gushed over one hand she spelled into the other the word *water*, first slowly, then rapidly. I stood still, my whole attention fixed upon the motions of her fingers. Suddenly I felt a misty consciousness as of something forgotten—a thrill of returning thought; and somehow the mystery of language was revealed to me. I knew then that "w-a-t-e-r" meant the wonderful cool something that was flowing over my hand. ...
>
> ... As we returned to the house every object which I touched seemed to quiver with life. That was because I saw everything with the strange, new sight that had come to me. (Keller 1905: 36.)

Moreover, if, indeed, it was '*milk*'—not '*water*'—that Helen was confusing with 'mug', then the well-house experience was a controlled experiment, filling the mug with *water* instead of milk.

Fourth, 'w-a-t-e-r' meant "the wonderful cool something that was flowing over my hand": 'W-a-t-e-r' was antecedently meaningless; "the wonderful cool something ..." was antecedently understood. As we have seen, the semantic relation is asymmetric; here we have the asymmetric equivalence of a definition (which is an intensional asymmetry).

Fifth, note that Helen did *not* say that 'water' meant H_2O: *Twin* Helen would have had the *same* experience, and 'water' would have meant exactly the same thing for her (modulo the essential indexical 'my'), viz., "the wonderful cool something that was flowing over my hand".

Finally, Helen's post–well-house experiences of seeing "everything with the strange, new sight" should be the eventual experience of Searle-in-the-room, who would then have *semantic* methods for doing things in addition to purely syntactic ones (cf. syntactic vs. semantic proofs in logic and math). Crucial to promoting semantics-as-correspondence to semantics-as-meaning—semantics-as-understanding—is that the semantic domain must be antecedently understood. This, as we shall see shortly, was crucial for Helen's progress.

9.3.2 Aftereffects.

Five days later, Annie Sullivan reports Helen replacing her own signs by the corresponding finger-spelled words as soon as she learns them (p. 257). Clearly, Helen had realized the advantages of this new, more efficient and expressive code for communication. Equally crucially, as she notes (p. 258), Helen *understood* what the finger-spelled words referred to *before* she was able to "utter" them. "The idea always precedes the word" (Annie Sullivan, 8 May 1887, p. 260). As Annie Sullivan noted later (pp. 291ff), Helen had her *own* signs for things before she had *words* for them, still using her signs when she had not yet learned the words (p. 261), so she was using two codes. She had several ways to communicate her ideas, preferring one (words), but using whatever was at hand (so to speak).

Two other observations that Annie Sullivan made are worth mentioning at this point. First, it was important for Helen to *generate* language, not merely to *understand* it, in order to help build her vocabulary (Annie Sullivan, 16 May 1887, pp. 262ff); interactive conversation is crucial (cf. §1.2.4 and Ch. 5).

Second,

> Language grows out of life, out of its needs and experiences. ... *Language* and *knowledge* are indissolubly connected; they are interdependent. Good work in language presupposes and depends on a real knowledge of things. As soon as Helen grasped the idea that everything had a *name*, and that by means of the manual alphabet these names could be transmitted from one to another, I proceeded to awaken her further interest in the *objects* whose names she learned to spell with such evident joy. *I never taught language for the* PURPOSE *of teaching it*; but invariably used language as a medium for the communication of *thought*; thus the learning of language was *coincident* with the aquisition of knowledge. In order to use language intelligently, one must have something to talk *about*, and having something to talk about is the result of having had experiences; no amount of language training will enable our little children to use language with ease and fluency unless they have something clearly in their minds which they wish to communicate, or unless we succeed in awakening in them a desire to know what is in the minds of others. (Annie Sullivan, in Keller 1905: 317.)

Bruner has observed much the same thing:

> So at the end of this first round of examining the simplest form of request—asking for objects—we are forced to a tentative conclusion. Language acquisition appears to be a by-product (and a vehicle) of culture transmission. Children learn to use a language initially (or its prelinguistic precursors) to get what they want, to play games, to stay connected with those on whom they are dependent. In so doing, they find the constraints that prevail in the culture around them embodied in their parents' restrictions and conventions. The engine that drives the enterprise is not language acquisition per se, but the need to get on with the demands of the culture. ... Children begin to use language ... not because they have a language-using capacity, but because they need to get things done by its use. Parents assist them in a like spirit: they want to help them become "civilized" human beings, not just speakers of the language. (Bruner 1983: 103–104.)

This is an insight that—beyond its evident importance for education *in general*—is of importance for *computational* natural-language understanding systems, too. It is not far from some of the insights of Dreyfus (1992). Whether it is something that can*not* be accomplished with computers remains, however, an empirical and open question, suggesting a clear direction for research.

Helen's language learning proceeded apace after the well house. The next month, she wrote her first letter to a friend (p. 123). Her vocabulary learning was cyclic and recursive—each new encounter with a word serving to clarify and enhance what she already knew (p. 40).

Words for abstract concepts (e.g., 'love', 'think')—concepts that could not be "shown", hence for which there was nothing apparent to associate them with—were harder, but not impossible, for her to learn (for the details, see Keller 1905: 40f, 300). In April 1887, she learned prepositions by direct experience—standing *on* a chair or *in* her wardrobe (Annie Sullivan's account, p. 279). Helen's own account of learning sentence structure is reminiscent of Russellian propositions: She would paste pieces of paper with words written on them onto the things they named: She would put her doll on the bed, the doll labeled 'doll', the bed labeled 'bed', with labels for 'is' and 'on' placed near the doll, on the bed; or she would put the label 'girl' on herself, the labels for 'is', 'in', and 'wardrobe' on the wardrobe, and then she would stand in the wardrobe, thus labeled.

Over a year later, by which time her language was of Turing Test quality, she would, nonetheless, use some not-yet-understood words in "parrotlike" fashion (Macy, in Keller 1905: 134) until she learned how to use them properly (until she learned their meaning?). These included "words of sound and vision which express ideas outside of her experience" (Macy, in Keller 1905: 134–135). I have argued that we do the same, with words like 'pregnant' used by a male (cf. Rapaport 1988: 116). Evidently, though, much more of Helen's knowledge is by description than is ours (cf. her description of a visit to an art gallery, Keller 1905: 200).

9.3.3 The Puzzle of the Well House.

But what really happened at the well house? The well-house association of 'water' with water was not different in kind from previous word–object associations that Helen had made and had used for communication. Annie Sullivan was not trying to teach Helen something new; she was merely trying to reinforce something she had more or less successfully taught her before. Various incidental experiences—Helen's mug/water-or-milk confusion, her memory of the spoken word 'wa-wa', and the perhaps unique "co-activation" of word and object (cf. Mayes 1991: 111)—no doubt contributed to making the well-house experience the significant event it was.

But Helen learned something she had not been taught. In her own and Annie Sullivan's words, she learned that things have "names". What exactly *did* she learn, and why was it so significant?

9.3.4 The Significance of the Well House.

One clear lesson that Helen learned on her own at the well house was, indeed, that things had names. But not just that, for merely knowing that 'w-a-t-e-r' or 'd-o-l-l' were the appropriate finger spellings to perform when in the presence of—or, more importantly, when *not* in the presence of, but desiring—water or a doll could be described as knowing that those things had names.[6] What Helen learned was that some things in the world, viz., finger spellings, were names of other things in the world. She learned the concept of a name, thereby learning a metalinguistic fact:[7] that her mental world was more than an associative network of concepts—it had a structure to it, in which some of the things in it (her internal representations of words) "named" others (her internal representations of objects, events, ideas, etc.). Roger Brown observed that "linguistic processes, in general tend to be invisible. The mind's eye seeks the meaning and notices the medium as little as the physical eye notices its own aqueous humor" (Brown 1973: 3). The well-house experience

[6] Possibly, 'd-o-l-l' means "Please give me my doll." Cf. "Please machine give cash" as the meaning of pushing a button; see §9.4.3.

[7] David Wilkins pointed out the metalinguistic nature of the well-house episode to me.

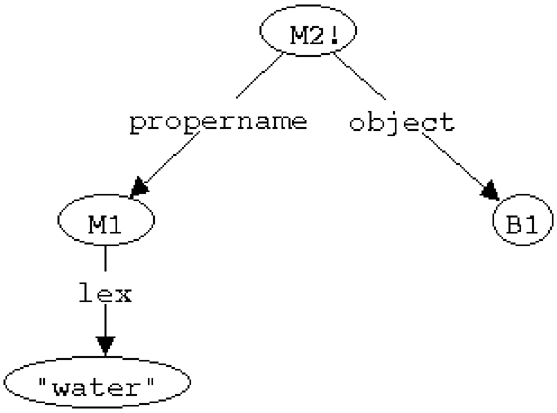


Figure 9.2: Before the well house. `M2` = This (`B1`) is (named) 'water'.

made one crucial linguistic process visible to Helen. In SNePS terms, Helen "promoted" a purely syntactic, inaccessible, labeled arc to a consciously accessible *node* (Figs. 9.2 and 9.3).

So, Helen learned more than how to *use* words or signs; she learned that certain things *were* words or signs. She learned more than how to "see through" a sign to its meaning; she learned how to see the sign *as a sign*.

9.4 TERRACE'S THEORY OF NAMES.

9.4.1 Introduction.

But why is it so significant to learn what a name is? A possible answer to this, consistent with Helen's post–well-house behavior, can be found in Herbert S. Terrace's theory of names. In this section, we'll look at Terrace's theory of why "naming" is important, whether his notion of "naming" is akin to Helen Keller's, the extent to which his theory is supported by more recent observations, and the relevance of all this to computational natural-language competence. We'll begin with a brief overview, and then look at his theory in detail.

9.4.2 Overview of T-Naming.

It will be both useful and convenient to distinguish Terrace's terms from, say, Helen's. So, I will refer to Terrace's theory of names and naming as the theory of 'T-names' and 'T-naming'.

In a letter to the editor of the *New York Review of Books*, Terrace summarizes his theory as follows:

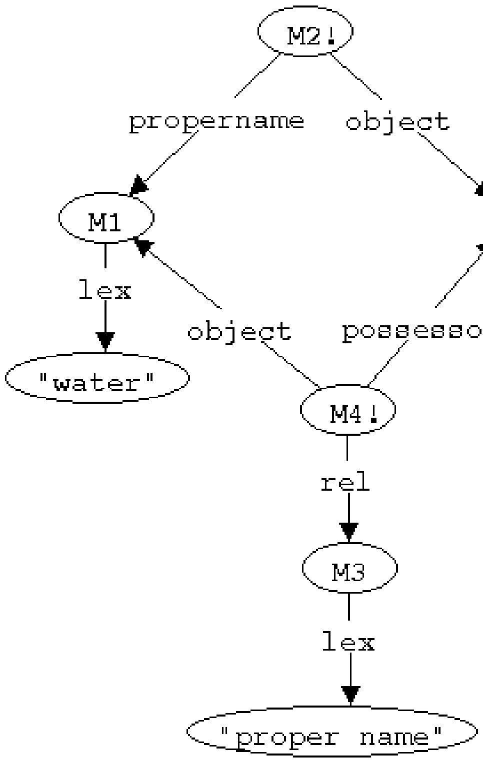


Figure 9.3: After the well house. `M4` = `M1` is `B1`'s *name*.

> Before speculating about the origins of grammar, it is prudent to ponder the origins of the *referential* use of individual words. Unlike apes, children use individual words to comment about objects for the sheer joy of communicating. Adults do not reward a child with a tree when she points to one and then says *tree* .... By contrast, there is no evidence that apes communicate *about* things. As Lord Zuckermann observes, apes use language not as "... a way of conversing, but a game associated with pleasurable reward." Although the origins of human language are unclear, one contributing factor must be the adaptive value of communicating meanings that cannot be expressed in a single word (e.g., *the large tree* or *the single-tusked elephant ate the large tree*). It appears, therefore, that the cognitive leap to language occurred in two stages: first, developing the lexical competence to use arbitrary symbols to refer to particular objects and events, and then the syntactic competence to combine and inflect those symbols systematically so as to create new meanings. (Terrace 1991: 53.)

The child's use of words for objects "for the sheer joy of communicating" clearly describes Helen's post–well-house behavior, and the game-like nature of language use by apes is reminiscent of Helen's pre–well-house use of language. So, *prima facie*, T-naming might well be what Helen learned to do at the well house. It was there that, by her metalinguistic discovery, she "develop[ed] the lexical competence to use arbitrary symbols to refer to particular objects and events".

Damasio observes that "the second stage [of concept formation] ... is that of generating names that are pertinent to the stimulus and are usable to narrate the primary display when inserted in appropriate grammatical structures" (Damasio 1989b: 25). Thus, I perceive an object, which causes neuronal activity representing its features and structure. These, in turn, are linked to other neuronal structures that "generate names", which allow me to communicate to you that I am thinking of an object and what it is.

Bruner makes a similar observation:

> In object request the principal task is to incorporate reference into request. When the child finally masters nominals, he [sic] need no longer depend upon the interpretive prowess of his mother or the deictic power of his indexical signalling. The demands of dealing with displaced reference in requesting objects provide an incentive. (Bruner 1983: 114.)

Having (T-)names gives one power. Helen, apparently, lacked this ability before the well house.

Of course, *pace* Terrace, one doesn't *create* new meanings. One creates new (combinations of) symbols that are able to be associated with things that one couldn't access linguistically before (cf. Elgin 1984 for a literary exploration of this theme). Zuckermann's way of putting this is better: "the additional adaptive value of joining lexical items in ways that multiplied the meanings that they can convey" (Zuckerman 1991: 53).

So, for Terrace, syntax is built on top of lexical semantics, as it seems to have been for Helen, too. Bruner concurs:

> ... requesting absent objects ... requires a degree of specification not needed when an object is within reach or sight. An object out of sight requires the use of nominals for easy specification. ...

> Remote or displaced requests began at the landmark age of fourteen months in both children. (Bruner 1983: 97–98.)

T-naming, as Figure 9.4 suggests, enables conversation—the exchange of *information*, distant or displaced in space and/or time, with no immediate, practical goals (other than, perhaps, to satisfy one's curiosity or to be sociable).

Let's now consider Terrace's claims in detail.

9.4.3 T-Naming.

In "In the Beginning Was the 'Name'" (1985), Terrace considers "The ability to refer with names" to be "perhaps" the most "fundamental" and "uniquely human skill" (p. 1011). This referential ability appears to be akin to symbol grounding. It is the link between word and world, reminiscent, perhaps, of the semantics of `LEX` nodes in SNePS. But reference, as Frege taught us, is not all there is to meaning. Is Terrace's notion of referring *Bedeutung*? Or *Sinn*? What would he say about the referring use of a name like 'unicorn' or 'Santa Claus'?

And if "no syntax without reference" is his rallying cry ("In the absence of referential naming, it is doubtful that syntax would have developed in human languages" (p. 1011)), what does that say about my theory of syntactic semantics? Of course, Terrace is concerned with *learning* syntax, not necessarily with having it already. Surely, however, learning syntax is ultimately more important than "hardwiring" it. How did Helen Keller learn to name? *What* did she learn? She learned the nature of the relationship between a name (i.e., a finger spelling) and an object. She learned the name of naming. So, is it possible for Cassie to learn to name? Given the network of Figure 9.2, we'd like her to be able to say, when asked, that 'water' is the *name* of `B2` (and not merely that this (`B2`) is called 'water'). And given the network in Figure 9.5, we'd like Cassie to be able to say, when asked, that red is a *property* of `B1` (and not merely that `B1` is red).[8] So, we would need to be able to have Cassie answer the following sort of questions: Given a propositional node asserting that some individual a has some property F, what is the (name of the) relationship between a and F?

Note that although Cassie can *use* 'water' as the name of `B2` (she can call `B2` 'water'), without a node explicitly asserting (naming) the relationship between `M3` and `B2`, she does not understand—*de dicto*—that 'water' is `B2`'s *name*. She has no theory of names or naming. Similarly, without a node explicitly asserting (naming) the relationship between `M1` and `B1` in Figure 9.5, she does not understand—*de dicto*—that red is a *property* of `B1`. She has no theory of properties.

Although these two situations are analogous, there is, perhaps, a slight advantage to the naming case over the property case. For in order for Cassie to *utter* something about `B1` or `B2`, she must use a word for it, whereas she need not use the word 'property' in uttering `M2`. So Cassie could, perhaps—this is a task for empirical investigation—recognize that there is a relationship between her concept of an object and the word she uses when she says something about it (between, that is, `B2` and `M3`). From this recognition, she could build a case frame that would link these nodes with a node for the relationship, which, if she asked us, we could tell her was called 'naming'. So

[8] Ann Deakin has pointed out in conversation that color is not a good example for Helen Keller! Perhaps taste would be better? On the other hand, for *Cassie*, color is currently more accessible than taste; cf. Lammens 1994.

she could build the case frame of Figure 9.3. In any event, let us suppose that this can be done. It is, it would seem, what Helen did.

Some of Terrace's claims about the linguistic abilities of apes are reminiscent of Helen's pre–well-house linguistic abilities:

> . . . even though apes can learn substantial vocabularies of arbitrary symbols, there is no a priori reason to regard such accomplishments as evidence of human linguistic competence. After all, dogs, rats, horse, and other animals can learn to produce arbitrary "words" to obtain specific rewards. (Terrace 1985: 1012.)

Helen "learn[ed] substantial vocabularies of arbitrary symbols", too. But what kind of learning is this? (Cf. Annie Sullivan's use of expressions like "Helen knew n words".) Given the context of Terrace's paper, it does not seem to mean that apes (or Helen) could link the arbitrary symbols to objects. And given Terrace's belief in the logical and chronological priority of naming over syntax, it does not seem to mean that the apes (or Helen) knew the syntactic roles of the symbols. Now, Helen, too, could "produce arbitrary 'words' to obtain specific rewards". So, by 'learning a symbol', Terrace must mean producing the symbol in return for a (not necessarily associated) reward, without any (semantic) linking of the symbol with the world.

> It would be just as erroneous to interpret the rote sequence of pecks [by a pigeon], red → green → blue → yellow, as a sentence meaning, *Please machine give grain*, as it would be to interpret the arbitrary sequence of digits that a person produces while operating a cash machine as a sentence meaning *Please machine give cash*. In sum, a rote sequence, however that sequence might be trained, is not necessarily a sentence. (Terrace 1985: 1014.)

This sounds like Helen's pre–well-house use of language. But *why* aren't those rote sequences sentences with those meanings? Granted, perhaps, they lack the same meaning (paraphrases exist, after all—propositions (meanings) can be *implemented* in numerous different ways, even within the same language). When I push a sequence of buttons on a cash machine, why *aren't* I telling (asking) it to give me a certain amount of money? Isn't that what the input–output encoding scheme amounts to? Granted, perhaps what the symbols mean to *me* is not what the symbols mean (if anything)[9] to the machine, but as long as we can communicate (so as to overcome misunderstandings), what's the difference?

A brief example might be instructive. In my university library, when I push the button for the elevator on the ground floor, the button lights up. I have come to learn empirically that if the light stays on when I release the button, it means that the elevator is not on the ground floor. If the light immediately goes off, it means that the elevator *is* on the ground floor and that in a few seconds the door will open. The light's going off is a symbol that I interpret to mean "the elevator is on the ground floor; the door will open shortly". What's going on here? The symbol in fact has two meanings. There is its meaning for me: my interpretation of it. I have determined this meaning empirically, and I could be wrong. If the light goes off and no elevator door opens within a few seconds (and it is in working order), I would have to revise my beliefs.

[9]But it *does* mean *something* to the machine—it has syntactic meaning (internal meaning).

There is also its meaning for the elevator system: the role that the light going off plays in the electrical network that controls the elevator. A study of the wiring diagram would reveal, let us suppose, that when the button is pushed, a circuit is closed that lights the button and a test is conducted to determine the location of the elevator. If the elevator is on the ground floor, the circuit is opened, turning off the light, and, a short time later, another circuit is closed, opening the door; else, the circuit remains closed, keeping the light on, and another circuit is closed, sending the elevator to the ground floor, where the light-circuit is opened, turning off the light, followed by the door opening as before. The meaning of the light's going off, then—its role in that network—is correlated with the elevator's being on the ground floor. From the elevator's point of view, so to speak, the only way the light going off would *not* mean that the elevator is on the ground floor would be if the elevator were broken.

One thing missing from such behavioral uses of language is the *intention* to communicate an idea by using a certain word (Terrace 1985: 1017). And although non-human animals who have been trained, behavioristically, to "use language" (which I place in scare quotes so as not to beg any questions about what it is they are actually doing) *seem* to communicate intentionally with each other. Terrace points out,

> That would presuppose not only that Jill [one of the pigeons] could discriminate each color from the others (when she clearly could) but that Jill also understood that (a) some arbitrary communicable symbol described color$_i$, (b) she sought to communicate to Jack [the other pigeon] that the color she saw was color$_i$, and (c) Jack would interpret Jill's message as Jill intended it. There is no evidence to support any of these suppositions. (Terrace 1985: 1016.)

This, of course, does not affect experiments in computational linguistics that endeavor to provide a mechanism (based on the theory of speech acts) for implementing intentions to speak (Bruce 1975; Cohen & Perrault 1979; Allen & Perrault 1980; Cohen & Levesque 1985, 1990; Grosz & Sidner 1986; Haller 1994, 1995). Indeed, that's one of the advantages a computer has over non-human animals: The latter might not have intentions to communicate, but the former can be given them.

But, according to Terrace, even if it could somehow be shown that a non-human animal *intended* to use a certain word to communicate the idea that it wanted a specific object, that would not suffice to show that it was using the word as a *name* for the object. Why? Presumably because it might believe that using that word is the best way to cause the listener to give it the desired object. Roughly, the animal might be ascribed the belief, "If I make such and such a noise [or use such and such a sign], my trainer will bring me one of those sweet, red, round things to eat, so I'll make that noise [or use that sign]".

So, for Terrace, *naming* is a very special activity:

> … the main function of such words [viz., the "use of a symbol as a name"] in the use of human language—[is] the transmission of information from one individual to another *for its own sake*. (Terrace 1985: 1016–1017; my italics.)

That is what I call "T-naming".

> … a child will utter a name of an object, person, color, and so on, simply to indicate

> that she or he knows that the object she or he is attending to has a name and also to communicate the fact that she or he has noticed that object ....
>
> ... In many instances, the child refers to the object in question spontaneously and shows no interest in obtaining it. The child not only appears to enjoy sharing information with his or her parent but also appears to derive intrinsic pleasure from the sheer act of naming. (Terrace 1985: 1017.)

A mere *associative* link between an arbitrary symbol and a specific object is not sufficient for a *semantic* link, according to Terrace. What is also needed is *intending* to use the symbol for the object *for no purpose other than to indicate that you are thinking of the object.*

So, one difference between Helen's pre- and post–well-house language might be that, *before*, she didn't have such intentions, and, *after*, she did. That is, *after*, she *intended* to refer to water by 'water'. To do this, she needed to be able to think and talk about the naming relationship. But is it really the case that she lacked that intention before? Although the evidence is at best unclear, I think she did have the intention, but not the name of the naming relationship, so that her intentions were often frustrated.

The Chinese Room (or Searle-in-the-room) *says* things but doesn't *mean* them, or so Searle would say. What, then, does it mean to mean something by a word? In Cassie's terms, it would be this: Cassie has a concept she wants to communicate to Oscar. She has a name for the concept. So, she utters that name, *assuming* that Oscar uses that word for the "same" (i.e, the corresponding) concept, and *intending* that Oscar will think of that concept—that that concept will be "activated" in Oscar's mind. As Terrace puts it,

> In most human discourse, a speaker who utters a name expects the listener to interpret the speaker's utterance as a reference to a jointly perceived (or imagined) object .... (Terrace 1985: 1017.)

"Jointly imagined" is where unicorns (not to mention Hob's and Nob's witch) come in. So, T-naming is more appropriately a relationship between a name and a mental concept (cf. Rapaport 1981). So, in the Chinese-Room Argument, what's missing from Searle's description (i.e., what's needed as one of the features of computational natural language competence in addition to those I cited in §1.2.4) is the *intention* to communicate. So one *could* argue that if Searle's Chinese Room is to be taken literally, then it *doesn't* understand, but that's because Searle hasn't fleshed out the full theory of computational natural-language competence; and if it's to be taken as schematic for whatever would be a full theory of computational natural-language competence, then he's wrong.

As both Terrace and Bruner (with his Language Acquisition Support System) point out—and as Annie Sullivan pointed out long before them—

> ... language draws upon certain kinds of nonlinguistic knowledge. For example, before learning to speak, an infant acquires a repertoire of instrumental behavior that allows her or him to manipulate and/or approach various objects. An infant also learns how to engage in various kinds of social interaction with her or his parents—for example, being able to look where the parent is looking or pointing. Eventually, the child learns to point to things that he or she would like the parent to notice. In short, the infant first masters a social and conceptual world onto which she or he can later map various kinds of linguistic expression. (Terrace 1985: 1018.)

So, *internal concepts* are learned first, via links between visual input and mental concepts (cf. `PIC` arcs: Rapaport 1988; Srihari & Rapaport 1989, 1990; Srihari 1991ab, 1993ab). Names are attached later (which suggests, by the way, that the `EXPRESSED-EXPRESSES` case frame is more appropriate than merely using `LEX` arcs).

Helen, thus, was able to learn language once she grasped the concept of naming—"the conventions of using symbols and words that do the work of referring" (Terrace 1985: 1021). Apes, according to Terrace, lack this ability (pp. 1021, 1023–1024). More specifically, "an ape does not naturally refer to an object to which it attends *solely for the purpose of noting that object to a conspecific*. ... [W]hatever referential skills an ape displays naturally seem to be in the service of some concrete end" (p. 1024, my italics). Helen's post–well-house interest in the names of objects seems to have been for its own sake. Her own pre–well-house signs were always "in the service of some concrete end".

But is Terrace's emphasis on intentional but non-purposive naming a reasonable one?

9.4.4 Critique of T-Naming.

Why does Terrace think that using a sign in order to get a reward is *not* using it linguistically (cf. Terrace 1985: 1016–1017)? What is so important about T-naming? And how do we know that apes don't have intentions when they "use language"? Finally, is there any evidence that apes *do* T-name?

To the charge that purposive-only use of names is not language, Terrace replies that it is simply a matter of fact that apes *don't* use signs except when they want something, whereas human children at 18 months *do*. So, at least, what apes do is preliminary to full human-like language use.

And that's what's important about T-naming, according to Terrace. Without it, grammatical syntax would not develop.

> ...when there is a desire simply to communicate information about a relationship between one object or action and another, about some attribute of an object, or about past or future events ... ungrammatical strings of words would not suffice—hence the functional value of syntax. (Terrace 1985: 1026.)

His argument seems to be this: If I want something that is present in the environment containing you and me, I can use a name to get you to give it to me. If a single name is insufficient (say, because there are two of them, and I want a specific one, e.g., the *large* banana), I can combine two (or more) names. But it doesn't really matter in what order I combine them, so grammar (beyond mere juxtaposition) is not necessary. If I don't know a name for the object, I can point to it. But for *absent* (displaced) objects, pointing won't work. And if I wanted to communicate about some *feature* of an (absent) object, grammar facilitates my communication: If I wanted to talk about the color of the banana, it becomes crucial *which* of the previously juxtaposed signs is the color-term and which the noun; grammar enters upon the scene. Note that it is highly unlikely that I *want* the color; rather, I just want you to know that I'm *thinking* about the color—that's T-naming.

As for intentions, we *don't* know that apes don't have them. I'd be willing to say that they do. We *can*, however, insure that a *computer's* use of language *is* intentional, by having the natural language competence program include speech-act/intention–action modules. Searle might object that that's just more syntax, not "real" intentions or desires. But what would be the difference? What is a "real" intention? Moreover, desires (and intentions) *can* be adequately incorporated. In the vocabulary-acquisition project, there will be times when the system, in order to settle on a definition, would need more information. That need—and why wouldn't it be a "real" need?—would prompt it to seek that information, to ask questions—and why wouldn't these be "real" desires for information or intentions to ask questions?

One might reply that such computational desires or intentions have no qualitative "feel" to them. Perhaps. Qualia, as I suggested, may best be seen as a feature of the implementing medium, not the Abstraction. So, of course, the computational desires and intentions *might* have a "feel", depending on their implementation. Or they might not. But why would a lack of "feeling" disqualify them as "real" desires or intentions?

As Figure 9.4 suggests, the thoughts, desires, and intentions of a language user that is not an ordinary human—an ape or a computer (or even, perhaps, a Helen Keller)—might be very different in kind from those of a normal human. They might very well depend on the language user's body and purposes. But they would be thoughts (and desires and intentions) nonetheless.

9.4.5 Can Apes Speak for Themselves?.

Is Terrace right about the inability of non-human primates to T-name? Several papers published *after* Terrace's deal with the issue specifically.

9.4.5.1 From representation to language.

Terrace claims that apes lack something that humans have that enables humans to have language. One researcher—Jacques Vauclair (1990)—is sympathetic to this, though it's not clear that that something is T-naming.

According to Vauclair (p. 312), both human and non-human primates have "basically similar ways of coding environmental stimuli in terms of cognitive organization"; i.e., they have "mental representations". But they do *not* share language. Why not? For Vauclair, it seems, it's partly by definition:

> *Representation* is an individual phenomenon by which an organism structures its knowledge with regard to its environment. This knowledge can take two basic forms: either reference to internal substitutes (e.g., indexes or images) or use of external substitutes (e.g., symbols, signals, or words).
>
> *Communication* is a social phenomenon of exchanges between two or more **conspecifics** who use a code of specific signals usually serving to meet common adaptive challenges (reproduction, feeding, protection) and promote cohesiveness of the group.
>
> (Vauclair 1990: 312; my boldface.)

Since apes and humans are not conspecifics, they cannot, *by definition*, communicate with each other. Even if that restriction is lifted, it is not clear whether T-naming is Vauclairian communication (unless, perhaps, as a by-product, it "promote[s] cohesiveness of the group"—perhaps that's the function of conversation; cf. Fig. 9.4).

> *Language* is conceived as a system that is both communicational and representational: It is grounded in social conversation that attributes to certain substitutes (called *signifiers*) the power to designate other substitutes (called *referents*).
>
> (Vauclair 1990: 313.)

So, apes and humans could never have a common language, because language is communicational (and, by definition, apes can't communicate with humans). But how does this definition of language make it *human*-specific? Perhaps it is the social-communication aspect (cf. Bruner's notion of "negotiation"). After all, apes and humans don't share a common "society".

The closest Vauclair gets to supporting Terrace's theory is in claiming that two of the marks of language are its ability to deal with displacement *in* space and time (i.e., things not in the presence of the speaker) and its ability to deal with what might be called displacement *from* space and time (i.e., dealing with non-existents) (Von Glasersfeld 1977, cited in Vauclair 1990: 316). T-naming, however, is logically *independent* of this. For one could, in principle, be able to refer to something displaced in (or from) space or time either if one wanted it or if one merely wanted to talk about it "for its own sake". And, clearly, one could be able to refer to something in one's current environment for either of those reasons.

So several issues are still open: *Do* non-human primates T-name? (Terrace, of course, says 'no'.) Do they use language to talk about displaced objects? One would expect Vauclair to delve into this. Instead, he locates the gap between ape and human language elsewhere:

> I am convinced that apes display the most sophisticated form of representation in the animal kingdom . . . , but this phenomenon is insufficient in itself to qualify for linguistic status. To go beyond the 1–1 correspondence between the sign and the actual perceptual situation, we need to introduce a third term. The relation between symbol and object is more than the simple correspondence between the two. Because the symbol is tied to a conception, we have a triangular connection among objects, symbols, and concepts: "It is the conceptions, not the things, that symbols directly mean" (Langer, quoted in von Glasersfeld, 1977). (Vauclair 1990: 320.)

Now, I am happy to agree with Langer, but it's not clear what that has to do with Vauclair's point. He *seems* here to be saying that what's missing is the concept: no concept, no language. Yet earlier he claimed that representation *required* concepts: Although 'concept' is not part of his *definition* of 'representation', on pp. 313ff he talks about "internal processing", "internal representation", "cognitive maps", "internal coding", and "internal substitutes". What are these if not concepts?

Later (p. 321), he locates the gap "in the emergence in humans of verbal language", but he is silent on what these emergent features are; perhaps it is T-naming. Or perhaps it is being intentional:

> The specificity of human language is above all of functional order. First, this system uses representative stimuli that allow the sender to know the status of the sent message, to control it, and to endow it with intentions. (Vauclair 1990: 321.)

Of course, as we have seen, this won't distinguish between human and *computer* use of language. Perhaps, however, this *was* something Helen Keller lacked before the well house.

The other thing that non-human primates lack is the social convention, the Brunerian negotiation of meaning (p. 322). This, however, seems irrelevant to T-naming. In any event, Helen Keller, arguably, had this *before* the well house, and computers certainly can have it (witness, e.g., the vocabulary-acquisition process).

9.4.5.2 Orangutan reference.

H. Lyn White Miles's work with the orangutan Chantek (1990) is suggestive of T-naming. Chantek was clearly capable of "displaced reference" (pp. 520–523), and four out of about 97 cited uses of *names* do not *appear* to involve wanting the object: making the signs (1) 'car' "as he passed [his] caregiver's car on a walk", (2) 'time' "when [his] caregiver looked at her pocket watch", (3) 'Coke drink' "*after* finishing his Coke" (my italics), and (4) 'time drink' "when [his] caregiver looked at her watch" (pp. 520–523). Each of these, however, could be interpreted otherwise: (1) Perhaps Chantek was tired of walking and wanted to ride in the car; (2) perhaps he wanted to know the time (though it's hard to believe that he had the appropriate concepts for understanding time) or perhaps 'time' was also his sign for the watch itself (we are not told); (3) perhaps he wanted another Coke to drink; (4) perhaps he was thirsty. It is hard to know when a naming act is a *T-naming*. Moreover, T-naming may be overly restrictive a criterion.

On the other hand, those who are more sympathetic than Terrace to the view that apes can use language tend to have criteria that are overly permissive. Consider Miles's three "elements" of "linguistic representation" (p. 524):

1. A sign must designate an element of the real world.
2. A shared cultural understanding about its meaning must exist.
3. The sign must be used intentionally to convey meaning.

The first element is surely too strong, since we can talk about nonexistents. Moreover, it would seem better to say that a sign must *be used by someone to* designate something (where 'something' is construed broadly, along Meinongian lines). The second element seems to rule out interspecies linguistic representation and, perhaps, computer language. On the other hand, in the various ape experiments, both subject and experimenter are using an artificial language, so they do have a shared cultural understanding, where the "culture" is that of the laboratory. Granted, the sign for Coke may have all sorts of connotations for the human but not the chimp. But that's no different from the fact that the word 'Coke' has all sorts of connotations for *me* but not *you*. The case of the computer is a bit easier, since we get to give it its cultural knowledge. Hence, insofar as the computer has a "mind" (i.e., a knowledge base, §1.2.5), we and it can have "shared cultural understanding", *pace* Dreyfus et al. (as long as we avoid Winston's Problem).

Helen Keller's pre–well-house uses of finger-spelled words seem in some cases to have designated in the sense of element (1) (e.g., some of her uses of 'cake' and 'doll'). Even her confused use of 'mug' and 'milk'/'water' might be taken to have designated the mug-plus-liquid complex. Clearly, before the well house, she could designate via her own signs. Arguably, her inability to clearly designate with finger spellings could be attributed to insufficiencies in her *shared* cultural understanding. She clearly shared in some cultural understanding—after all, she was a human, living with other humans. But, of course, she was blind and deaf, hence cut off from much that the rest of us share without even realizing it. Finally, though she used her own signs intentionally to convey meaning, most of her pre–well-house use of finger spellings was no doubt mere mimickry.

Again, Miles's criteria for referential use of words or signs is weaker than Terrace's:

> first, that signs can be used to indicate an object in the environment; second, that signs are not totally context dependent; third, that signs have relevant semantic domains or realms of meaning; fourth, that signs can be used to refer to objects or events that are not present. (Miles 1990: 524.)

(I take the third criterion to mean that there is a systematic correlation between sign and referent.) All of these are necessary—but not sufficient—for T-naming. One of the essential aspects of T-naming is that there be no desire to *have* the object named—no ulterior motive.

However, Chantek showed some behavior that seems to be part of T-naming when he would show his caregivers some object (pp. 524–525). Since he already had the object, it would seem that he had no other purpose for showing it than to get his caregivers to understand that he was thinking about it. This behavior, when combined with displaced reference, surely lays the groundwork for eventual T-naming.

Is T-naming a significant mark either of human language development in particular or of language development *simpliciter*? Granted that Helen Keller exhibited it after (and apparently *only* after) the well house, it would seem that it is significant for humans (or, at least, for her). And if Chantek either could easily have exhibited, or in fact did exhibit (on occasion), T-naming, it might *not* be unique to human language. It certainly makes for more sophisticated use of language (the ability to tell stories, the ability to fabricate), and it does make language learning easier. Yet there's an awful lot of linguistic behavior that apes such as Chantek are capable of that makes one wonder why Terrace requires that in order to T-name, the language user must *not* want the object. Chantek, for instance, learned labels for things he wanted, displayed displacement reference for things he wanted, and used language to deceive in order to get something (pp. 526–529). And Chantek, apparently, was capable of a metalinguistic achievement that, again, could underlie eventual T-naming:

> By transferring the total shape of the sign, including configuration and movement, to another means of expression **[WHAT?]**, he showed that he understood that the sign was an abstract representation in which the composite elements stood for something else. (Miles 1990: 530.)

Indeed, some of the beginnings of what looks like T-naming can be seen in the following passages:

> The second stage of development, that of subjective representation ... ranged from 2 years to almost $4\frac{1}{2}$ years of age .... In this stage, Chantek used his signs as symbolic representations, but his perspective remained subjective. He gave the first evidence of displacement ... and developed proximal pointing, which indicated that he had mental representations. ... He elaborated his deception and pretend play .... He showed

> evidence of planning through mental representations and signed to himself about objects not present. ... For the first time he also used signs in his deceptions. (Miles 1990: 534–535.)

> The third stage, nascent perspective taking, ranged from about $4\frac{1}{2}$ years to over 8 years of age, during which his vocabulary increased to 140 signs .... Chantek's representations became more objective and moved toward perspective taking, the ability to utilize the point of view of the other. ... Most important, he was able to take the perspective of the other by getting the caregiver's attention and directing the caregiver's eye gaze before he began to sign.
>
> It was at this point that he invented his own signs. ... He clearly understood that signs were representational labels, and he immediately offered his hands to be molded when he wanted to know the name of an object. (Miles 1990: 535.)

How reminiscent of Helen Keller's post–well-house behavior, whether or not it is T-naming!

9.4.5.3 Against T-naming.

Two arguments can be mounted *against* the significance of T-naming. The first, due to Patricia Marks Greenfield and E. Sue Savage–Rumbaugh (1990), is based on possible biases on the part of researchers. Terrace's claim that apes don't T-name is apparently supported by evidence such as that "Kanzi [a pygmy chimpanzee] had a much smaller proportion of indicatives to statements (4%) in comparison with requests (96%), than would be normal for a human child" (Greenfield & Savage–Rumbaugh 1990: 568). But, as Greenfield and Savage–Rumbaugh point out, an alternative explanation is that this is an artifact of their artificial, human-controlled environment, in which they *must* request things. By contrast, "In the wild, a given animal might *state* his planned activity, rather than *requesting* it" (p. 568). They suggest that if we studied human language development without the assumption that children will eventually succeed in learning language, we might *not* ascribe T-naming to them at the analogous developmental stage at which we deny it to apes (p. 571).

The second, perhaps weaker, argument against T-naming focusses on just what it is that a speaker intends to communicate. T-naming certainly involves a desire to communicate—but to communicate what? For Terrace, it is the desire to communicate that the speaker is thinking of a distal object. The speaker is playing a sort of "guess what I'm thinking about" game, using a word that means *what he or she is thinking about*. But that notion—what the speaker is thinking about—is ambiguous between the actual object (a *de re* interpretation) and the concept in the speaker's mind (a *de dicto* interpretation). However, since the speaker can be thinking of an object that doesn't exist, or a proposition that may lack a truth value, the *de re* interpretation fails. Only the *de dicto* interpretation can be consistently maintained in all cases (Rapaport 1976, 1978, 1981, 1985/1986, 1986a). As a consequence, *all* uses of names turn out to be T-naming.

9.5 WHO (OR WHAT) CAN HAVE REPRESENTATIONS?

In the introduction to his book *Models* (1979), Wartofsky returns to what I labeled theses 2 and 5 in §2.6.2, arguing that for one thing to be a representation of another, something must do the representing, and, moreover, the representer must be human. Although I agree that a cognitive agent can take one thing as a representation or model of another, it is worth noting that this is facilitated to the extent that there are agent-independent correspondences that "afford" the use of one thing as a model of another. More significantly, I see no reason for reserving to humans alone the ability to make or use such correspondences. We looked at the need for, and role of, a cognitive agent (cf. §2.7.1). My concern here will be whether that agent needs to be human.

That humans *can* and do make and use representations seems obvious. That this ability is *essential to cognition* is, perhaps, less so. (For some recent arguments (with which I disagree) that it is *not* essential, see Brooks 1991ab.) That it is *unique to human cognition* also requires justification. In the quotation from Wartofsky that opened §1.2, I omitted a crucial word; the full statement is this:

> the crucial feature of *human* cognitive practice [is] ... the ability to make representations." (Wartofsky 1979: xiii, my italics.)

Thus, for Wartofsky, the representational ability is both essential and unique to human cognition. I certainly agree that it is essential—that is one of my main themes (though I will not deal with the apparent counterclaims of Brooks). But why need it be unique? We can grant, with Wartofsky (1979: xviii), that "Strictly speaking, one may say that there is no human knowledge without representation;" but we need not go on to agree that "more radically still, ... there is no knowledge without representation [...] that knowledge, suitably defined, is a distinctively human achievement, and that it is to be qualitatively distinguished from animal intelligence and learning ...." And similarly for machine intelligence and learning.

Granted, *human* knowledge, intelligence, and learning is "natural" whereas AI is "artificial". But if there is no knowledge without representation, then where there *is* representation, there can be knowledge. So the question is: Can non-human animals and machines make and use representations? Wartofsky is willing to say 'yes', but only metaphorically. He wants to reserve "the term *representation* ... to be used in a way which demarcates human from non-human animal consciousness and activity" since "animals neither make nor use representations in their conscious activity, (though they may use what *we* make as, or take to be representations)." The use of 'representation' "in any talk about 'internal representation' in non-human animals, or indeed, any talk of representation in machines, e.g. in computers, or by means of mechanical or electronic or chemical repreduction, is likewise metaphorical, anthropomorphic and parasitical for its meaning upon the human activity of representing" (Wartofsky 1979: xix).

There are a number of issues here. Perhaps the easiest to deal with is the claim that non-human animals don't make or use representations. This appears to be empirically false. C. R. Gallistel has provided overwhelming evidence in favor of the claim that a wide variety of non-human animals, including various kinds of birds, rats, bees, monkeys, ants, and fish are capable of making and using a variety of kinds of representations of space, time, number, rate, and social relations (Gallistel 1989, 1990ab). For Gallistel, a representation is a "functional isomorphism", where by 'isomorphism' he means the kinds of correspondences we have been looking at, and by 'functional' he means that the animal must actually *use* the representation to "generate valid anticipations of events and relations in the represented system" (Gallistel 1990b: 2)—to make internal predictions of external states of affairs. This is in fact a much narrower notion than what Wartofsky has in mind. So if non-human animals make and use such functional isomorphisms, surely they are capable of making and using representations in Wartofsky's more liberal sense.

What about the case of computers? Well, this is what my entire essay is about! Clearly, though, suitably-programmed computers can and do make and use representations. Some, perhaps—such as certain connectionist or other machine-learning systems—even make their own, without human help (cf. Winston 1975). The issue is whether such representations are anything more than "mere" metaphor. I think they are, for two reasons. First, if something is "taken" as a representation, then, by definition, it really *is* a representation. Second, in some cases—and representation is one of them—the metaphor is only apparent. I elaborated on both of these reasons in Chapter 7. For now, consider the following anticipation by Wartofsky of some recent debates in philosophy of mind (cf. Dretske 1985; Laymon 1988; Rapaport 1988, 1990, 1993b; D. Cole 1991, Hauser 1993):

> Such conceptual and terminological *caveats* are not intended ... to condemn the extended usage of 'knowing' and 'representing' which pervades our talk about ... machine computation and reproduction. Rather, the intention is to recognize that such usages are metaphorical, rather than mistakes; and mistaken only when it is forgotten that they are metaphorical and anthropomorphic. In fact, an enlightened anthropomorphism is perhaps the most powerful heuristic framework for inquiry into ... computational process. For example, it is useful and fruitful ... to think of machines as 'adding', 'solving equations' or 'comparing and assessing probabilities'. I would not even want to propose that we redescribe this by saying "machines do not add ... but do something *like* what we do when *we* add ...". It is not necessary that computers be said to do 'something like what we do', to avoid the anthropomorphism, because they may be 'doing' *nothing at all* 'like what we do' in these contexts. It is enough that we can use the *model* of 'adding' ... *to represent to ourselves*, and to understand better what it is the machines are 'doing', or how they operate (since a machine may be 'doing' nothing at all, but simply moving from one state to another, in accordance with a program). (Wartofsky 1979: xix–xx.)

Wartofsky's "enlightened anthropomorphism" seems to be Dennett's "intentional stance" (1971). But, I would argue, machine-state transitions *are* "doings": If there is a level of description at which a computer can be *said to be* adding, then it *is* adding, and its machine-state transitions are *how* it does it (this is spelled out in Rapaport 1990).

Wartofsky does, however, have an interesting insight on the methodology of computational cognitive science, one that suggests that much of the debate over that methodology is misplaced. He tells us that

> [c]onversely, it may be that we come to understand aspects of what **we** do, in computing or solving problems, by representing it in a model of **machine** operations, stripped of all (or nearly all) anthropomorphic metaphor. But again, the model need not be taken in any sense as an account of what we *do*, when we compute or reason. (Wartofsky 1979: xx; italics in original, boldface added.)

That is, even if there is a computational model of some human cognitive process, it does not follow that the *way* we humans do it is computational.

Does this make sense? Let P be some cognitive process that we humans do (say, reasoning or perceiving or natural-language understanding). Suppose A_P is an algorithm with the same input–output behavior that we exhibit when we do P. Thus, A_P is a computational model of P. Does it follow that P itself is a computational process? That is, does it follow that *we* do P computationally? I'm afraid not. After all, we might do it by magic; or by some mysterious process that can only be described in stimulus–response terms; or by some highly analog, "messy" physical process. I think, in fact, this is unlikely; but it is surely logically possible.

But let f be some function-in-extension, that is, some appropriate set of ordered pairs. Then the *function* f is recursive (or "computable") if and only if *there is* an *algorithm* A_f with the same input–output "behavior" as f—that is, an algorithm that computes f. There are *two* things: f and A_f. Similarly, P is computable if and only if *there is* some algorithm A_P. But P itself need not be algorithmic. There is, thus, a distinction between "being comput*able*" and "*being* a computational process": P and f can be comput*able without* being computational processes; A_P and A_f *are* computational processes. Thus, just because there may be a way to do P by means of water pipes and valves (to use one of Searle's examples), it doesn't follow that *humans* do P that way; after all, we use *neurons*. So, clearly, to find out how *humans* do P, we will need a *neurophysiological* theory. *But*—and this is a *big* 'but'—first, without a computational theory to *guide* us (even one made of water pipes and valves), our neurophysiological theory will be hard, if not impossible, to develop (this is the point made early on by Chomsky (1968)). Second, cognitive science ought not to be limited to studying how *humans* do P. The central question of cognitive science (as opposed to, say, cognitive psychology) should be the Kantian question: *How is cognition possible?* How is cognition possible *independent* of the (physical) medium in which it is implemented? And here, a computational theory (if one exists) suffices. Such is computational cognitive science considered as computational *philosophy* (cf. Rapaport 1986a, Shapiro 1992a).

9.6 BRUNER AGAIN.

Let us return to the theme of this chapter, from which we have digressed somewhat: What was the significance of the well-house episode? Bruner's *Child's Talk* (1983) offers some ideas that are relevant (cf. §5.3, above).

As we've noted several times, negotiation is crucial to understanding language. Two interlocutors must endeavor to align the concepts that each finds or builds when a word is used. Equally, one often has to *merge* two of one's own concepts or to *split* one into two (cf. Maida & Shapiro 1982). So, one thing that was significant about Helen's experience at the well house was that two of *her* concepts merged or were equated: *her* concept of water (previously linked to 'wa-wa') and *her* concept of *Annie Sullivan's* concept of water. Prior to the well house, Annie Sullivan *thought* that these were merged ideas, but, in fact, they weren't.

Moreover, the well house itself played a significant role:

> ... a key feature of human referring acts ... [is that] [t]hey are highly context sensitive or deictic. Parties to a referring act infer its referent from an *utterance* in a *context*.

> ... John Lyons argues that deixis is the source of reference, that "locating in context" rather than simply "tagging" is the heart of reference .... (Bruner 1983: 69–70.)

(On the importance of deixis for natural-language understanding, cf. Bruder et al. 1986; Rapaport, Segal, Shapiro, Zubin, Bruder, Duchan, Almeida et al. 1989; Rapaport, Segal, Shapiro, Zubin, Bruder, Duchan & Mark 1989; Duchan et al. 1995.) Helen's experience was significant because the context was extremely clear and simple: water in one hand, 'water' in the other.[10]

> One might reasonably expect to find, then, that the acquisition of referring procedures is heavily dependent on the "arranging" and simplifying of contexts by the adult to assure that deictic demands be manageable for the child. (Bruner 1983: 70.)

9.7 CONCLUDING REMARKS ON HELEN KELLER.

The story of Helen Keller is a fascinating one. Every teacher, not only of language skills, ought to read Annie Sullivan's letters and reports on her teaching methods. But Helen was an amazing pupil. (One wonders what would have become of her had she *not* been blind and deaf!) Consider, for example, the large number of syntactic systems with which she was familiar: finger spelling (the manual alphabet, *tactually* understood), lip reading (again, tactually understood), the typewriter, three varieties of Braille, the Roman alphabet (again, tactually understood), oral speech (her own),[11] Morse code, English, French, German, Latin, and Greek (and probably the Greek alphabet in some form).

Now, Searle-in-the-room also knows a syntactic system—squiggles—which are known to others as Chinese writing. The task for Searle-in-the-room is to get "beyond" the syntax to ... what? To ideas? To objects? In general, of course, his task is to get to what the squiggles *mean*. How? Well, clearly, the more squiggles, the better. Note that *much* of Helen Keller's learning was *book*-learning, which is purely syntactic (cf. Keller 1905: 308, 318; but cf. p. 317). But also Searle-in-the-room needs more experiences, even if only self-bodily ones. But, ultimately, *all* such experiences are internal(ly represented), just as are (the experiences of) the squiggles.

Ditto for Helen Keller. When she was able to *organize* all her internal symbols such that some were names for others (and some of the "others" were directly linked to her experiences), she began to get beyond the syntax to the meanings (cf. Keller 1905: 169). The organizing principle was discovered at the well house.

9.8 SUMMARY.

In this book, I hope to have built a holistic web of ideas that cohere. Some reinforce others; some provide inferential support for others. The fundamental idea is that understanding is recursive—we understand one domain in terms of an antecedently understood one, and, in the base case, we understand some domains syntactically (in terms of themselves). In syntactically understood

[10] Actually, as we saw, there was a mug in the water hand, but it seems to have been ignored.

[11] Helen's knowledge of speech is also akin to the Chinese Room: She had a "syntactic" knowledge of speech, since she couldn't hear herself. Cf. Keller 1905: 327.

domains, some elements are understood in terms of others. In the case of language, linguistic elements are understood in terms of non-linguistic ("conceptual") yet internal elements. Thus can semantics arise from syntax. And that is how natural-language understanding, by human or computer, is possible.

Figure 9.4:

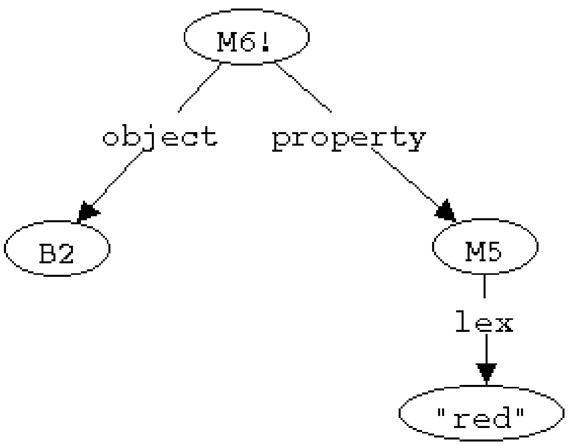
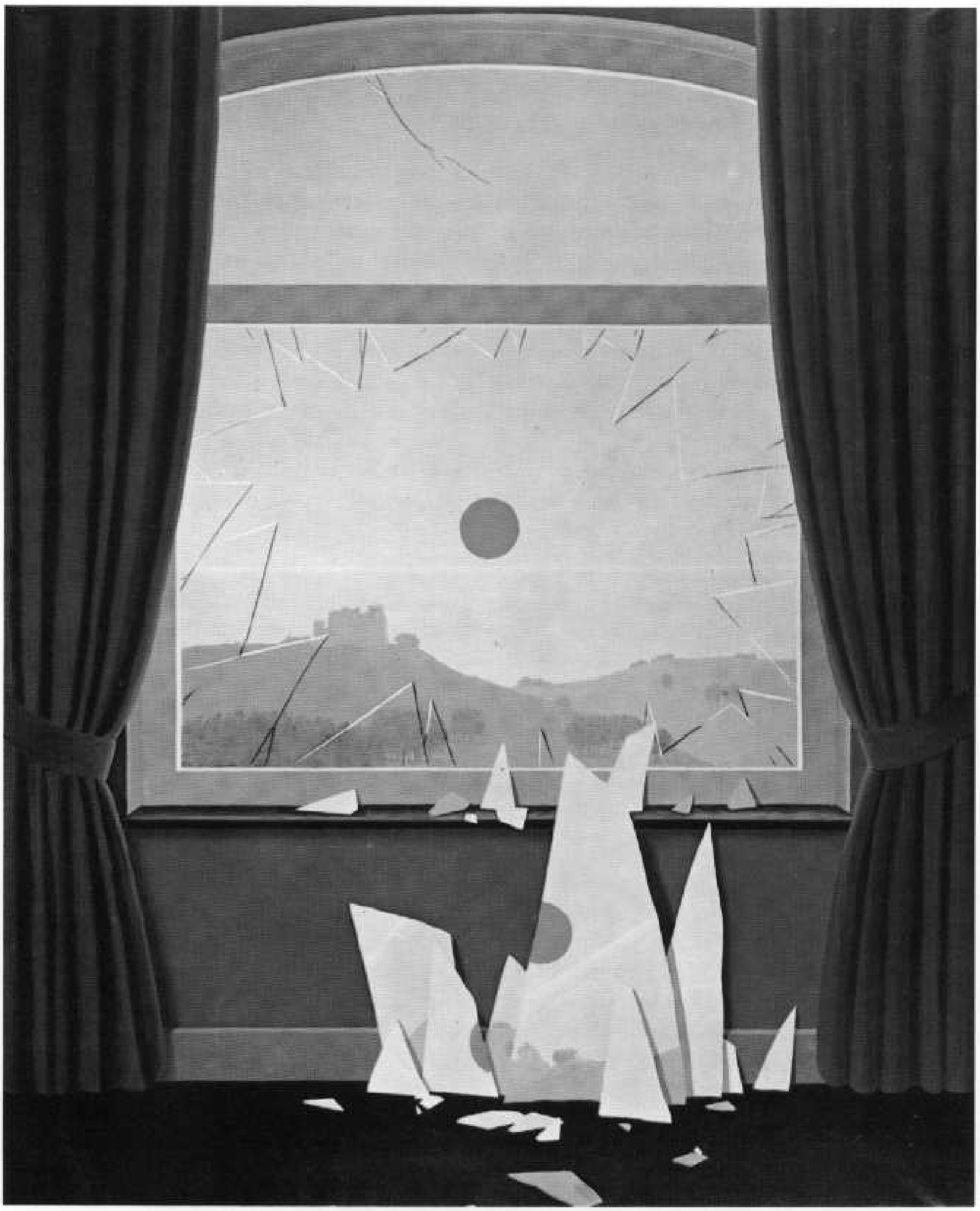
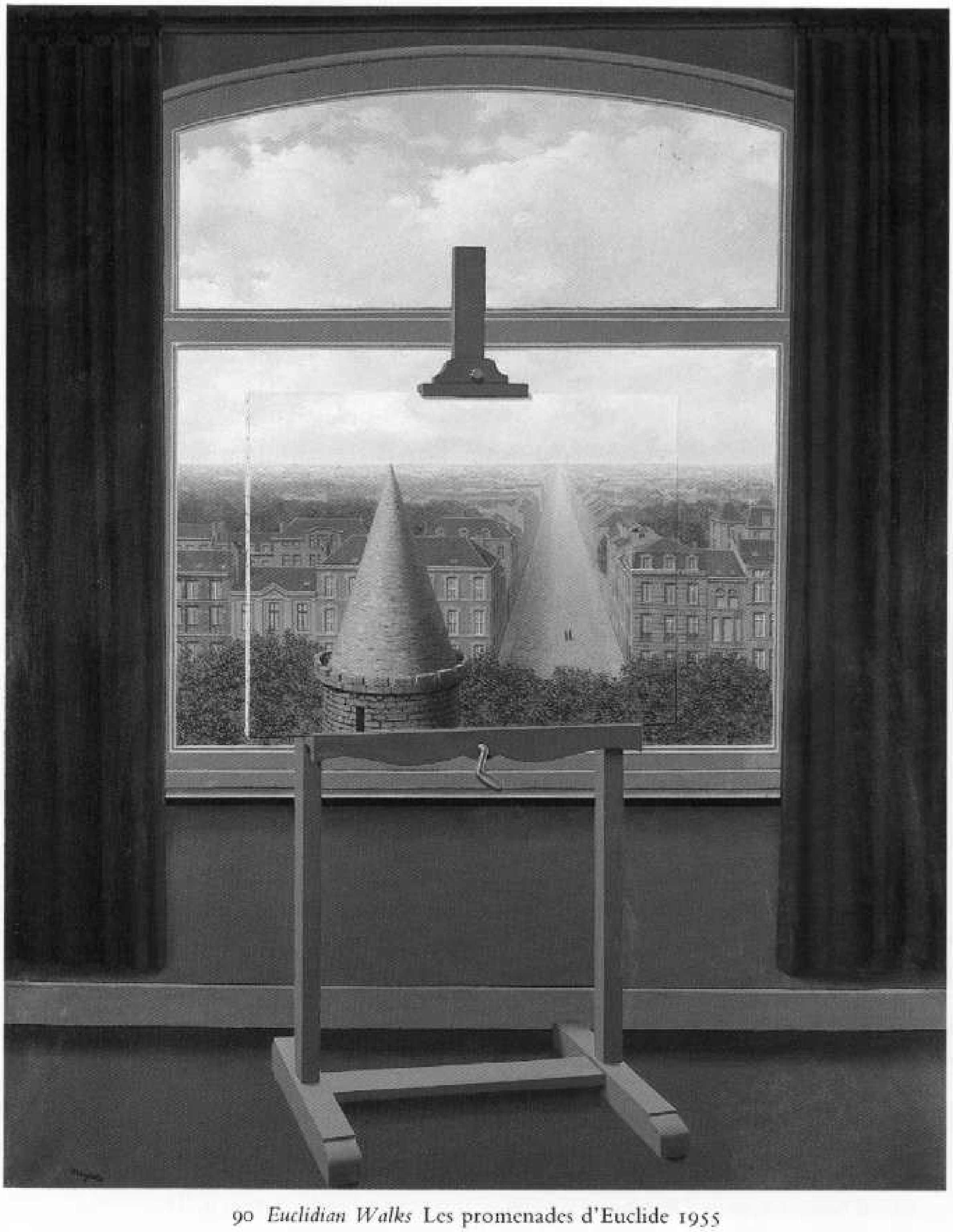


Figure 9.5:



79 *Evening Falls* Le soir qui tombe 1964



90 *Euclidian Walks* Les promenades d'Euclide 1955

Chapter 10

REFERENCES.

1. Ackerman, Diane (1989), "Penguins," *The New Yorker* (10 July 1989): 000–000.

2. Ali, Syed S. (1994), "A 'Natural Logic' For Natural Language Processing and Knowledge Representation," *Technical Report 94-01* (Buffalo: SUNY Buffalo Department of Computer Science).

3. Ali, Syed S. (1995), "ANALOG: A Knowledge Representation System for Natural Language Processing," in E. A. Yfantis (ed.), *Intelligent Systems: 3rd Golden West International Conference* (Dordrecht, Holland: Kluwer Academic Publishers): 327–332.

4. Ali, Syed S., & Shapiro, Stuart C. (1993), "Natural Language Processing Using a Propositional Semantic Network with Structured Variables," *Minds and Machines* 3: 421–451.

5. Allaire, Edwin B. (1963), "Bare Particulars," *Philosophical Studies* 14; reprinted in Michael J. Loux (ed.), *Universals and Particulars* (Garden City, NY: Anchor Books, 1970): 235–244.

6. Allaire, Edwin B. (1965), "Another Look at Bare Particulars," *Philosophical Studies* 16; reprinted in Michael J. Loux (ed.), *Universals and Particulars* (Garden City, NY: Anchor Books, 1970): 250–257.

7. Allen, James F., & Perrault, C. Raymond (1980), "Analyzing Intentions in Utterance," *Artificial Intelligence* 15: 143–178; reprinted in Barbara J. Grosz, Karen Sparck Jones, & Bonnie L. Webber (eds.), *Readings in Natural Language Processing* (Los Altos, CA: Morgan Kaufmann, 1986): 441–458.

8. Almeida, Michael J. (1987), "Reasoning about the Temporal Structure of Narratives," *Technical Report 87-10* (Buffalo: SUNY Buffalo Department of Computer Science).

9. Almeida, Michael J. (1995), "Time in Narratives," in Judith F. Duchan, Gail A. Bruder, & Lynne E. Hewitt (eds.), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates).

10. Almeida, Michael J., & Shapiro, Stuart C. (1983), "Reasoning about the Temporal Structure of Narrative Texts," *Proceedings of the 5th Annual Conference of the Cognitive Science Society (University of Rochester)* (Hillsdale, NJ. Lawrence Erlbaum Associates).

11. Altmann, Stuart A. (1989), "The Monkey and the Fig: A Socratic Dialogue on Evolutionary Themes," *American Scientist* 77: 256–263.

12. Alvarez, Celso (3 May 1990), article 6387, from [sp299-ad@violet.berkeley.edu] (a.k.a. [celso@athena.berkeley.edu], on [sci.lang] electronic bulletin board.

13. Angier, Natalie (1992), "Odor Receptors Discovered in Sperm Cells," *The New York Times* (30 January 1992): A19.

14. Anthony, Michael V. (1991), "Fodor and Pylyshyn on Connectionism," *Minds and Machines* 1: 321–341.

15. Apostel, Leo (1961), "Towards the Formal Study of Models in the Non-Formal Sciences," in Hans Freudenthal (ed.), *The Concept and the Role of the Model in Mathematics and Natural and Social Sciences* (Dordrecht, Holland: D. Reidel): 1–37.

16. Araki, Kenji, & Momouchi, Yoshio (1990), "Learning of Semantic Concept in Copular Sentence" (in Japanese), *IPSJ SIG Reports*, Vol. 90, No. 77.

17. Asher, Nicholas (1986), "Belief in Discourse Representation Theory," *Journal of Philosophical Logic* 15: 127–189.

18. Baker, Nicholson (1994), "Discards," *The New Yorker* (4 April 1994): 64–86.

19. Baker, Robert (1967), "Particulars: Bare, Naked, and Nude," *Noûs* 1: 211–212.

20. Barker, Clive (1987), *Weaveworld* (New York: Poseidon).

21. Barwise, Jon, & Etchemendy, John (1989), "Model-Theoretic Semantics," in Michael I. Posner (ed.), *Foundations of Cognitive Science* (Cambridge, MA: MIT Press): 207–243.

22. Barwise, Jon, & Perry, John (1983), *Situations and Attitudes* (Cambridge, MA: MIT Press).

23. Berwick, Robert C. (1979), "Learning Structural Descriptions of Grammar Rules from Examples," *Proceedings of the 6th International Joint Conference on Artificial Intelligence (IJCAI-79, Tokyo)* (Los Altos, CA: William Kaufmann): 56–58.

24. Berwick, Robert C. (1980), "Computational Analogues of Constraints on Grammars: A Model of Syntactic Acquisition," *Proceedings of the 18th Annual Conference of the Association for Computational Linguistics (Toronto)* (Morristown, NJ: Association for Computational Linguistics): 49–53.

25. Biermann, Alan 1990, *Great Ideas in Computer Science: A Gentle Introduction* (Cambridge, MA: MIT Press).

26. Blakeslee, Sandra (1993), "Human Nose May Hold An Additional Organ For a Real Sixth Sense," *New York Times* (7 September): C3.

27. Blish, James (1970), *Spock Must Die!* (New York: Bantam Books).

28. Borges, Jorge Luis (1981), "Partial Enchantments of the Quixote," in Emir Rodriguez Monegal & Alastair Reid (eds.), *Borges, A Reader; A Selection from the Writings of Jorge Luis Borges* (New York: E. P. Dutton): 232–235.

29. Brachman, Ronald J., & Schmolze, James G. (1985), "An Overview of the KL-ONE Knowledge Representation System," *Cognitive Science* 9: 171–216.

30. Brentano, Franz (1874), "The Distinction between Mental and Physical Phenomena," trans. D. B. Terrell, in Roderick M. Chisholm (ed.), *Realism and the Background of Phenomenology* (New York: Free Press, 1960): 39–61.

31. Bringsjord, Selmer (1992), *What Robots Can and Can't Be* (Dordrecht, Holland: Kluwer Academic Publishers).

32. Brooks, Rodney A. (1991a), "Intelligence without Representation," *Artificial Intelligence* 47: 139–159.

33. Brooks, Rodney A. (1991b), "Intelligence without Reason," *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91; Sydney, Australia)* (San Mateo, CA: Morgan Kaufmann): 569–595.

34. Brown, Roger (1973), *A First Language: The Early Stages* (Cambridge, MA: Harvard University Press).

35. Bruce, Bertram C. (1975), "Generation as a Social Action," *Theoretical Issues in Natural Language Processing–1* (Morristown, NJ: Association for Computational Linguistics): 64–67; reprinted in Barbara J. Grosz, Karen Sparck Jones, & Bonnie L. Webber (eds.), *Readings in Natural Language Processing* (Los Altos, CA: Morgan Kaufmann, 1986): 419–422.

36. Bruder, Gail A.; Duchan, Judith F.; Rapaport, William J.; Segal, Erwin M.; Shapiro, Stuart C.; & Zubin, David A. (1986), "Deictic Centers in Narrative: An Interdisciplinary Cognitive-Science Project," *Technical Report 86-20* (Buffalo: SUNY Buffalo Department of Computer Science).

37. Bruner, Jerome (1983), *Child's Talk: Learning to Use Language* (New York: W. W. Norton).

38. Calvino, Italo (1986), "Why Read the Classics?" *The New York Review of Books* (9 October 1986): 19–20.

39. Canby, Vincent (1994), "Yet Another Evolution for 'Singin' in the Rain'," *New York Times* (25 September).

40. Carnap, Rudolf (1928), *The Logical Structure of the World*, Rolf A. George (trans.), (Berkeley: University of California Press, 1967).

41. Carnap, Rudolf (1956), *Meaning and Necessity: A Study in Semantics and Modal Logic, 2nd edition* (Chicago: University of Chicago Press).

42. Carroll, Lewis (1895), "What the Tortoise Said to Achilles," *Mind* 4: 278–280; reprinted in *Mind* 104: 691–693, and in William Warren Bartley, III (ed.), *Lewis Carroll's Symbolic Logic* (New York: Clarkson N. Potter, 1977): 431–434. See also commentary, pp. 466–475.

43. Castañeda, Hector-Neri (1972), "Thinking and the Structure of the World," *Philosophia* 4 (1974) 3–40; reprinted in 1975 in *Critica* 6 (1972) 43–86.

44. Castañeda, Hector-Neri (1975), "Identity and Sameness," *Philosophia* 5: 121–150.

45. Castañeda, Hector-Neri (1975b), "Individuation and Non-Identity: A New Look," *American Philosophical Quarterly* 12: 131–140.

46. Castañeda, Hector-Neri (1977), "Perception, Belief, and the Structure of Physical Objects and Consciousness," *Synthese* 35: 285–351.

47. Castañeda, Hector-Neri (1979), "Fiction and Reality: Their Fundamental Connections; An Essay on the Ontology of Total Experience," *Poetics* 8: 31–62.

48. Castañeda, Hector-Neri (1980), "Reference, Reality, and Perceptual Fields," *Proceedings and Addresses of the American Philosophical Association* 53: 763–823.

49. Castañeda, Hector-Neri (1989a), *Thinking, Language, and Experience* (Minneapolis: University of Minnesota Press).

50. Castañeda, Hector-Neri (1989b), "Objects, Identity, and Sameness," *Topoi* Supp. Vol. 4, pp. 31–64.

51. Castañeda, Hector-Neri (1989c), "The Reflexivity of Self-Consciousness: Sameness/Identity, Data for Artificial Intelligence," *Philosophical Topics* 17: 27–58.

52. Castañeda, Hector-Neri (1989d), "The Reflexivity of Self-Consciousness: Sameness/Identity, Data for Artificial Intelligence," *Philosophical Topics* 17: 27–58.

53. Chang, Chen-Chung, & Keisler, H. Jerome (1973), *Model Theory* (Amsterdam: North–Holland).

54. Cho, Kah-Kyung (1992), "Re-thinking Intentionality" (in Japanese), in Y. Nitta (ed.), *Tasha-no Genshogaku (Phenomenology of the Other)* (Hokuto Publishing Co.).

55. Chomsky, Noam (1968), *Language and Mind* (New York: Harcourt, Brace & World).

56. Churchland, Patricia S., & Sejnowski, Terrence J. (1992), *The Computational Brain* (Cambridge, MA: MIT Press).

57. Clancey, William J. (1991), Book Review of Israel Rosenfield, *The Invention of Memory*, in *Artificial Intelligence* 50: 241–284.

58. Coffa, J. Alberto (1991), *The Semantic Tradition from Kant to Carnap: To the Vienna Station* (Cambridge, UK: Cambridge University Press).

59. Cohen, Philip R., & Levesque, Hector J. (1985), "Speech Acts and Rationality," *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (University of Chicago)* (Morristown, NJ: Association for Computational Linguistics): 49–60.

60. Cohen, Philip R., & Levesque, Hector J. (1990), "Rational Interaction as the Basis for Communication," in Philip R. Cohen, Jerry Morgan, & Martha E. Pollack (eds.), *Intentions in Communication* (Cambridge, MA: MIT Press): 221–256.

61. Cohen, Philip R., & Perrault, C. Raymond (1979), "Elements of a Plan-Based Theory of Speech Acts," *Cognitive Science* 3: 177–212; reprinted in Barbara J. Grosz, Karen Sparck Jones, & Bonnie L. Webber (eds.), *Readings in Natural Language Processing* (Los Altos, CA: Morgan Kaufmann, 1986): 423–440.

62. Colby, Kenneth M., & Smith, David Canfield (1969), "Dialogues between Humans and an Artificial Belief System," *Proceedings of the [1st] International Joint Conference on Artificial Intelligence (IJCAI-69; Washington, DC)* (Los Altos, CA: William Kaufmann, Inc.): 319–324.

63. Cole, David (1990), "Functionalism and Inverted Spectra," *Synthese* 82: 207–222.

64. Cole, David (1991), "Artificial Intelligence and Personal Identity," *Synthese* 88: 399–417.

65. Cole, Jerald D. (1991), "WHILE Loops and the Analogy of the Single Stroke Engine," *SIGCSE Bulletin* 23.3 (September 1991): 20–22.

66. Copeland, Jack (1993), *Artificial Intelligence: A Philosophical Introduction* (Oxford: Blackwell Publishers).

67. Corless, Robert M. (1992), "Continued Fractions and Chaos," *American Mathematical Monthly* 99: 203–215.

68. Coughlin, Ellen K. (1991), "A Professor Champions Distinct Culture of Deaf People," *Chronicle of Higher Education* (2 October 1991): A5.

69. Cullingford, Richard (1981), "SAM," in Roger C. Schank & Christopher K. Riesbeck, *Inside Computer Understanding: Five Programs Plus Miniatures* (Hillsdale, NJ: Lawrence Erlbaum Associates): 75–119.

70. Cushing, James T. (1991), "Quantum Theory and Explanatory Discourse: Endgame for Understanding?", *Philosophy of Science* 58: 337–358.

71. Damasio, Antonio R. (1989a), "Time-Locked Multiregional Retroactivation: A Systems-Level Proposal for the Neural Substraates of Recall and Recognition," *Cognition* 33: 25–62.

72. Damasio, Antonio R. (1989b), "Concepts in the Brain," in "Forum: What is a Concept?", *Mind and Language* 4: 24–27.

73. Davidson, Donald (1970), "Mental Events," in L. Foster & J. W. Swanson (eds.), *Experience and Theory* (Amherst: University of Massachusetts Press): 79–101.

74. Dennett, Daniel C. (1971), "Intentional Systems," *Journal of Philosophy* 68: 87–106; reprinted in Daniel C. Dennett, *Brainstorms* (Montgomery, VT: Bradford Books): 3–22.

75. Dennett, Daniel C. (1978), "Why You Can't Make a Computer that Feels Pain," *Synthese* 38: **00–00**; reprinted in Daniel C. Dennett, *Brainstorms* (Montgomery, VT: Bradford Books): 190–229.

76. Dennett, Daniel C. (1982), "Beyond Belief," in Andrew Woodfield (ed.), *Thought and Object: Essays on Intentionality* (Oxford: Clarendon Press): xvi–95.

77. Dipert, Randall R. (1990), "Complexity and Models of Minds: A Simple, Hilbertian Argument that Strong AI is Doomed" (Fredonia, NY: SUNY Fredonia Department of Philosophy). Paper presented at the Computers and Philosophy conference, Stanford, CA.

78. Dretske, Fred I. (1981), *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press).

79. Dretske, Fred I. (1985), "Machines and the Mental," *Proceedings and Addresses of the American Philosophical Association* 59: 23–33.

80. Dreyfus, Hubert L. (1992), *What Computers* Still *Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press).

81. Duhamel, Georges (1931), *Les Jumeaux de Vallangoujard* [*The Twins of Vallangoujard*], M. E. Storer (ed.) (Boston: D. C. Heath, 1940).

82. Duchan, Judith F.; Bruder, Gail A.; & Hewitt, Lynne (eds.) (1995), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates).

83. Durfee, Edmund H. (1992), "What Your Computer Really Needs to Know, You Learned in Kindergarten," *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92; San Jose, CA)* (Menlo Park CA: AAAI Press/MIT Press): 858–864.

84.

85. Eco, Umberto (1982), "On the Impossibility of Drawing a Map of the Empire on a Scale of 1 to 1," William Weaver (trans.), in Umberto Eco, *How to Travel with a Salmon and Other Essays* (New York: Harcourt Brace, 1994): 95–106. Eco, Umberto (1988), "On Truth. A Fiction," in Umberto Eco, Marco Santambrogio, & Patrizia Violi (eds.), *Meaning and Mental Re23presentations* (Bloomington: Indiana University Press): 41–59.

86. Ehrlich, Karen (1995), "Automatic Vocabulary Expansion through Narrative Context," *Technical Report 95-09* (Buffalo: SUNY Buffalo Department of Computer Science).

87. Ehrlich, Karen, & Rapaport, William J. (1992), "Automatic Acquisition of Word Meanings from Natural-Language Contexts," *Technical Report 92-03* (Buffalo: SUNY Buffalo Center for Cognitive Science, July 1992).

88. Ehrlich, Karen, & Rapaport, William J. (1993), "Vocabulary Expansion through Natural-Language Context," *Proceedings of the 8th Annual University at Buffalo Graduate Conference on Computer Science* (Buffalo: SUNY Buffalo Department of Computer Science): 78–84.

89. Ehrlich, Karen, & Rapaport, William J. (1995), "A Computational Theory of Vocabulary Expansion: Project Proposal," *Technical Report 95-15* (Buffalo: SUNY Buffalo Department of Computer Science) and *Technical Report 95-08* (Buffalo: SUNY Buffalo Center for Cognitive Science).

90. Elgin, Suzett Haden (1984), *Native Tongue* (New York: DAW Books).

91. Fetzer, James H. (1988), "Program Verification: The Very Idea," *Communications of the Association for Computing Machinery* 31: 1048–1063.

92. Fetzer, James H. (1991), "Philosophical Aspects of Program Verification," *Minds and Machines* 1: 197–216.

93. Field, Hartry 1977

94. Flanagan, Owen J. (1984), *The Science of Mind* (Cambridge, MA: MIT Press).

95. Fodor, Jerry A. (1968), *Psychological Explanation: An Introduction to the Philosophy of Psychology* (New York: Random House).

96. Fodor, Jerry A. (1975), *The Language of Thought* (New York: Thomas Y. Crowell Co.).

97. Fodor, Jerry A. (1980), "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," *Behavioral and Brain Sciences* 3: 63–109.

98. Fodor, Jerry A. (1981), "The Mind–Body Problem," *Scientific American* (January): 114–123.

99. Fodor, Jerry A. (1983), *The Modularity of Mind: An Essay in Faculty Psychology* (Cambridge, MA: MIT Press).

100. Fodor, Jerry, & Lepore, Ernest (1991), "Why Meaning (Probably) Isn't Conceptual Role," *Mind and Language* 6: 328–343.

101. Fodor, Jerry, & Lepore, Ernest (1992), *Holism: A Shopper's Guide* (Cambridge, MA: Basil Blackwell).

102. Frank, Robert J. (1990), cited in "Marginalia" column, *Chronicle of Higher Education* (1 August 1990): A2.

103. Frege, Gottlob (1892), "On Sense and Reference," Max Black (trans.), in Peter Geach & Max Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege* (Oxford: Basil Blackwell, 1970): 56–78.

104. Gablik, Suzi (1985), *Magritte* (New York: Thames & Hudson).

105. Galbraith, Mary, & Rapaport, William J. (eds.) (1991), "Where Does *I* Come From? Subjectivity and the Debate over Computational Cognitive Science," *Technical Report 91-01* (Buffalo: SUNY Buffalo Center for Cognitive Science, May 1991).

106. Gallistel 1989

107. Gallistel 1990a

108. Gallistel 1990b

109. Gobbi 84

110. Goguen, J. A.; Thatcher, J. W.; & Wagner, E. G. (1978), "An Initial Algebra Approach to the Specification, Correctness, and Implementation of Abstract Data Types," in Raymond T. Yeh (ed.), *Current Trends in Programming Methodology*, Vol. IV: *Data Structuring* (Englewood Cliffs, NJ: Prentice–Hall): 80–149.

111. Govindaraju, Venu; Sher, David B.; Srihari, Rohini K.; & Srihari, Sargur N. (1989), "Locating Human Faces in Newspaper Photographs," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR-89, San Diego)*: 278–285.

112. Govindaraju, Venu, & Srihari, Rohini K. (1990), "Recognizing Faces in a News Photo Database," *Advanced Imaging* 5: 22–26.

113. Gracia, Jorge J. E. (1990), "Texts and Their Interpretation," *Review of Metaphysics* 43: 495–542.

114. Graubard, Stephen R. (ed.) (1988), "Artificial Intelligence," special issue of *Daedalus*, Vol. 117, No. 1 (Winter 1988); reprinted as *The Artificial Intelligence Debate: False Starts, Real Foundations* (Cambridge, MA: MIT Press, 1988).

115. Greenfield, Patricia Marks, & Savage-Rumbaugh, E. Sue (1990), "Grammatical Combination in *Pan Paniscus*: Processes of Learning and Invention in the Evolution and Development of Language," in S. T. Parker & K. R. Gibson (eds.), *"Language" and Intellect in Monkeys and Apes* (Cambridge, UK: Cambridge University Press): 540–577.

116. Grosz, Barbara J., & Sidner, Candace L. (1986), "Attention, Intentions, and the Structure of Discourse," *Computational Linguistics* 12: 175–204.

117. Guttag, John V.; Horowitz, Ellis; & Musser, David R. (1978), "The Design of Data Type Specifications," in Raymond T. Yeh (ed.), *Current Trends in Programming Methodology*, Vol. IV: *Data Structuring* (Englewood Cliffs, NJ: Prentice-Hall): 60–79.

118. Haller, Susan M. (1993a), "Planning for Intentions with Rhetorical Relations," *Proceedings of a Workshop on Intentionality and Structure in Discourse Relations (Ohio State University)* (Morristown, NJ: Association for Computational Linguistics): 23–26.

119. Haller, Susan M. (1993b), "Collaboration in an Interactive Model of Plan Explanation," *Proceedings of a AAAI Fall Symposium on Human-Computer Collaboration: Reconciling Theory, Synthesizing Practice, Technical Report FS93-05* (Menlo Park, CA: AAAI).

120. Haller, Susan M. (1994), "Interactive Generation of Plan Descriptions and Justifications," *Technical Report 94-40* (Buffalo, NY: SUNY Buffalo Department of Computer Science).

121. Haller, Susan M. (1995), "Planning Text for Interactive Plan Explanations," in E. A. Yfantis (ed.), *Intelligent Systems: 3rd Golden West International Conference* (Dordrecht, Holland: Kluwer Academic Publishers): 61–67.

122. Hardt, Shoshana L. (1992), "Conceptual Dependency," in Stuart C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence, 2nd Edition* (New York: John Wiley & Sons): 259–265.

123. Harman, Gilbert (1974), "Meaning and Semantics," in Milton K. Munitz & Peter K. Unger (eds.), *Semantics and Philosophy* (New York: New York University Press): 1–16.

124. Harman, Gilbert (1975), "Language, Thought, and Communication," in Keith Gunderson (ed.), *Minnesota Studies in the Philosophy of Science; Vol. 7: Language, Mind, and Knowledge* (Minneapolis: University of Minnesota Press): 270–298.

125. Harman, Gilbert (1982), "Conceptual Role Semantics," *Notre Dame Journal of Formal Logic* 23: 242–256.

126. Harman, Gilbert (1987), "(Nonsolipsistic) Conceptual Role Semantics," in Ernest Lepore (ed.), *New Directions in Semantics* (London: Academic Press): 55–81.

127. Harman, Gilbert (1988), "Wide Functionalism," in Stephen Schiffer & Ssusan Steele (eds.), *Cognition and Representation* (Boulder, CO: Westview Press): 11–18.

128. Harnad, Stevan (1990), "The Symbol Grounding Problem," *Physica D* 42: 335–346.

129. Harris, Roy (1987), *Reading Saussure: A Critical Commentary on the* Cours de Linguistique Générale (La Salle, IL: Open Court).

130. Haugeland, John (1983), "Phenomenal Causes," *Southern Journal of Philosophy*, Vol. 22 (Supp.): 63–70.

131. Haugeland, John (1985), *Artificial Intelligence: The Very Idea* (Cambridge, MA: MIT Press).

132. Hauser, Larry (1993), "Why Isn't My Pocket Calculator a Thinking Thing?," *Minds and Machines* 3: 3–10.

133. Hayes, John P. (1988), *Computer Architecture and Organization, 2nd edition* (New York: McGraw-Hill).

134. Hayes, Patrick J. (1990), "Searle's Chinese Room," abstract of presentation to the Society for Philosophy and Psychology Symposium on Searle's Chinese Room and Workshop on Symbol Grounding, University of Maryland (electronic mail posting from Stevan Harnad, 6 June 1990).

135. Hedrick, C. (1976), "Learning Production Systems from Examples," *Artificial Intelligence* 7: 21–49.

136. Hexmoor, Henry (1995), "Representing and Learning Routine Activities," *Technical Report 95-xx* (Buffalo: SUNY Buffalo Department of Computer Science).

137. Hexmoor, Henry; Lammens, Johan; Caicedo, Guido; & Shapiro, Stuart C. (1993a), "Behavior Based AI, Cognitive Processes, and Emergent Behaviors in Autonomous Agents,j " in G. Rzevski, J. Pastor, & R. Adey (eds.), *Applications of AI in Engineering VIII, Vol. 2, Applications and Techniques* (CITY: CMI/Elsevier): 447–461.

138. Hexmoor, Henry; Lammens, Johan; & Shapiro, Stuart C. (1993b), "Embodiment in GLAIR: A Grounded Layered Architecture with Integrated Reasoning," in Douglas D. Dankell II (ed.), *Florida AI Research Symposium* (Ft. Lauderdale: FLAIRS): 325–329; also *Technical Report 93-10* (Buffalo: SUNY Buffalo Department of Computer Science).

139. Hexmoor, Henry; Lammens, Johan; & Shapiro, Stuart C. (1993c), "An Autonomous Agent Architecture for Integrating 'Unconscious' and 'Conscious' ,Reasoned Behaviors," in EDITOR, *Computer Architectures for Machine Perception* (New Orleans: PUB?): 000–000.

140. Higginbotham, James (1985), "On Semantics," reprinted in Ernest Lepore (ed.), *New Directions in Semantics* (London: Academic Press, 1987): 1–54.

141. Higginbotham, James (1989), "Elucidations of Meaning," *Linguistics and Philosophy* 12: 465–517.

142. Hill, Robin K. (1994), "Issues of Semantics in a Semantic-Network Representation of Belief," *Technical Report 94-11* (Buffalo: SUNY Buffalo Department of Computer Science).

143. Hill, Robin K. (1995), "Non-Well-Founded Set Theory and the Circular Semantics of Semantic Networks," in E. A. Yfantis (ed.), *Intelligent Systems: 3rd Golden West International Conference* (Dordrecht, Holland: Kluwer Academic Publishers): 375–386.

144. Hirst, Graeme (1989), "Ontological Assumptions in Knowledge Representation," *Proceedings of the 1st International Conference on Principles of Knowledge Representation and Reasoning (KR-89, Toronto)* (San Mateo, CA: Morgan Kaufmann): 157–169.

145. Hirst, Graeme (1991), "Existence Assumptions in Knowledge Representation," *Artificial Intelligence* 49: 199–242.

146. Hirst, Graeme; McRoy, Susan; Heeman, Peter; Edmonds, Philip; & Horton, Diane (1993), "Repairing Conversational Misunderstandings and Non-Understandings," *International Symposium on Spoken Dialogue (Waseda University, Tokyo).*

147. Ho, Tin Kam (1990), "A Multiple Classifier Approach to Visual Word Recognition," Ph.D. dissertation proposal (Buffalo: SUNY Buffalo Department of Computer Science, August 1990).

148. Hofstadter, Douglas R. (1981), "A Coffeehouse Conversation on the Turing Test," *Scientific American* (May): 15–36; reprinted with Reflections (by Daniel C. Dennett) in Douglas R. Hofstadter & Daniel C. Dennett (eds.), *The Mind's I: Fantasies and Reflections on Self and Soul* (New York: Basic Books, 1981): 68–95; and reprinted with Post Scriptum in Douglas R. Hofstadter, *Metamagical Themas: Questing for the Essence of Mind and Pattern* (New York: Basic Books, 1985): 492–525.

149. Hoyle, Fred (1987), *The Black Cloud* (New York: Harper & Row).

150. Jackson, Frank (1986), "What Mary Didn't Know," *Journal of Philosophy* 83: 291–295.

151. Jahren, Neal (1990), "Can Semantics Be Syntactic?", *Synthese* 82: 309–328.

152. James, William (1893), *The Principles of Psychology*, Vol. I (New York: Henry Holt).

153. Jardine, Nicholas (1973), "Model-Theoretic Semantics and Natural Language," in Edward L. Keenan (ed.), *Formal Semantics of Natural Language* (Cambridge, UK: Cambridge University Press, 1975): 219–240.

154. Jennings, Richard C. (1985), "Translation, Interpretation and Understanding," paper read at the American Philosophical Association Eastern Division (Washington, DC); abstract, *Proceedings and Addresses of the American Philosophical Association* 59 (1985) 345–346.

155. Johnson, George (1990), "New Mind, No Clothes," *The Sciences* (July/August 1990): 45–49.

156. Johnson, Mark (1987), *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason* (Chicago: University of Chicago Press).

157. Johnson-Laird, Philip N. (1983), *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Cambridge, MA: Harvard University Press).

158. Johnson-Laird, Philip N. (1988), *The Computer and the Mind: An Introduction to Cognitive Science* (Cambridge, MA: Harvard University Press).

159. Kamp, Hans (1983), "Situations in Discourse without Time or Questions" (Austin: University of Texas Center for Cognitive Science).

160. Kamp, Hans (1984), "A Theory of Truth and Semantic Representation," in J. Groenendijk, T. M. V. Janssen, & M. Stokhof (eds.), *Truth, Interpretation, and Information* (Dordrecht, Holland: Foris): 1–41.

161. Kamp, Hans (1985), "Context, Thought and Communication," *Proceedings of the Aristotelian Society*, N.S. 85 (1984/1985): 239–261.

162. Kamp, Hans (1988), "Comments on Stalnaker," in Robert H. Grimm & Daniel D. Merrill (eds.), *Contents of Thought* (Tucson: University of Arizona Press), pp. 156–181.

163. Kamp, Hans, & Reyle, Uwe (1993), *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory* (Dordrecht, Holland: Kluwer Academic Publishers).

164. Kay, Alan C. (1991), "Computers, Networks and Education," *Scientific American* Vol. 265, No. 3 (September 1991) 138–148.

165. Keller, Helen (1905), *The Story of My Life* (Garden City, NY: Doubleday, 1954).

166. Kim, Jaegwon (1974), "Noncausal Connections," *Noûs* 8: 41–52.

167. Kim, Jaegwon (1978), "Supervienience and Nomological Incommensurables," *American Philosophical Quarterly* 15: 149–156.

168. Kim, Jaegwon (1979), "Causality, Identity, and Supervenience in the Mind–Body Problem," in Peter A. French, Theodore E. Uehling, Jr., & Howard K. Wettstein (eds.), *Midwest Studies in Philosophy*, Vol. 4: *Studies in Metaphysics* (Minneapolis: University of Minnesota Press): 31-49.

169. Kim, Jaegwon (1982), "Psychophysical Supervenience," *Philosophical Studies* 41: 51–70.

170. Kim, Jaegwon (1983), "Supervenience and Supervenient Causation," *Southern Journal of Philosophy* 22, Supplement, pp. 45–56.

171. Kirsh, David (1991), "Foundations of AI: The Big Issues," *Artificial Intelligence* 47: 3–30.

172. Knight, Kevin (1989a), "A Gentle Introduction to Subsymbolic Computation: Connectionism for the A.I. Researcher," *Technical Report CMU-CS-89-150* (Pittsburgh: Carnegie Mellon University, School of Computer Science).

173. Knight, Kevin (1989b), "Unification: A Multidisciplinary Survey," *ACM Computing Surveys* 21: 93–124.

174. Knight, Kevin (1990), "Connectionist Ideas and Algorithms," *Communications of the ACM* Vol. 33, No. 11, pp. 59–74.

175. Kosslyn, Stephen M. (1981), "The Medium and the Message in Mental Imagery: A Theory," in Ned Block (ed.), *Imagery* (Cambridge, MA: MIT Press): 207–244.

176. Kosslyn, Stephen M.; Pinker, Steven; Smith, George E.; & Schwartz, Steven P. (1981), "On the Demystification of Mental Imagery," in Ned Block (ed.), *Imagery* (Cambridge, MA: MIT Press): 121–150.

177. Kumar, Deepak (1993a), "A Unified Model of Acting and Inference," in Jay F. Nunamaker, Jr., & Ralph H. Sprague, Jr. (eds.), *Proceedings of the 26th Hawaii International Conference on System Sciences*, Vol. III: *Decision Support Systems and Knowledge-Based Systems* (Los Alamitos, CA: IEEE Computer Society Press): 483–492.

178. Kumar, Deepak (1993b), "An AI Architecture Based on Message Passing," in James Geller (ed.), *Proceedings of the 1993 AAAI Spring Symposium on Innovative Applications of Massively Parallel Architectures* (Menlo Park, CA: AAAI Press): 127–131.

179. Kumar, Deepak (1993c), "Rational Engines for BDI Architectures," in Amy Lansky (ed.), *Proceedings of the 1993 AAAI Spring Symposium on Foundations of Automated Planning* (Menlo Park, CA: AAAI Press): 78–82.

180. Kumar, Deepak (1994a), "The SNePS BDI Architecture," *Journal of Decision Support Systems* ??: 000–000.

181. Kumar, Deepak (1994b), "From Beliefs and Goals to Intentions and Actions: An Amalgamated Model of Acting and Inference," *Technical Report 94-04* (Buffalo: SUNY Buffalo Department of Computer Science).

182. Kumar, Deepak, & Shapiro, Stuart C. (1993), "Deductive Efficiency, Belief Revision, and Acting," *Journal of Experimental and Theoretical Artificial Intelligence* 5: 167–177.

183. Kumar, Deepak, & Shapiro, Stuart C. (1995), "The OK BDI Architecture," in E. A. Yfantis (ed.), *Intelligent Systems: 3rd Golden West International Conference* (Dordrecht, Holland: Kluwer Academic Publishers): 307–317.

184. Kundera, Milan (1978), *The Book of Laughter and Forgetting*, Michael Henry Heim (trans.) (New York: Penguin Books).

185. Lakoff, George (1987), *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* (Chicago: University of Chicago Press).

186. Lakoff, George cogssoc

187. Lakoff, George, & Johnson, Mark (1980), *Metaphors We Live By* (Chicago: University of Chicago Press).

188. Lammens 49

189. Lammens, Johan (1994), "A Computational Model of Color Perception and Color Naming," *Technical Report 94-26* (Buffalo: SUNY Buffalo Department of Computer Science).

190. Lammens, Johan; Hexmoor, Henry; & Shapiro, Stuart C. (1995), "Of Elephants and Men," *NATO ASI Series on Biology and Technology of Intelligent Autonomous Agents* (CITY: Springer-Verlag): 000–000.

191. Landman, Fred (1986), "Pegs and Alecs," in Joseph Y. Halpern (ed.), *Theoretical Aspects of Reasoning About Knowledge* (Los Altos, CA: Morgan Kaufmann): 45–61.

192. Langley, P. (1980), "A Production System Model of First Language Acquisition," *Proceedings of the 8th International Conference on Computational Linguistics (COLING–80, Tokyo)*: 183–189).

193. Langley, P. (1982), "Language Acquisition through Error Recovery," *Cognition and Brain Theory* 5: 211–255.

194. Laymon, Ron (1988), "Some Computers Can Add (Even if the IBM 1620 Couldn't): Defending ENIAC's Accumulators against Dretske," *Behaviorism* 16: 1–16.

195. Lewis, David (1972), "General Semantics," in Donald Davidson & Gilbert Harman (eds.), *Semantics of Natural Language* (Dordrecht, Holland: D. Reidel): 169-218.

196. Lewis, David (1975), "Languages and Language," in Keith Gunderson (ed.), *Minnesota Studies in the Philosophy of Science; Vol. 7: Language, Mind, and Knowledge* (Minneapolis: University of Minnesota Press): 3–35.

197. Leiber, Justin (1980), *Beyond Rejection* (New York: Ballantine Books).

198. Leiber, Justin (1991), *An Invitation to Cognitive Science* (Cambridge, MA, and Oxford: Basil Blackwell).

199. Lenat, Douglas B. (1995), "CYC, WordNet, and EDR: Critiques and Responses—Lenat on WordNet and EDR," *Communications of the ACM* 38.11 (November) 45–46.

200. Lenat, Douglas B., & Feigenbaum, Edward A. (1991), "On the Thresholds of Knowledge," *Artificial Intelligence* 47: 185–250.

201. Lepore, Ernest, & Loewer, Barry (1981), "Translation Semantics," *Synthese* 48: 121–133.

202. Levesque, Hector J. (1986), "Making Believers out of Computers," *Artificial Intelligence* 30: 81–108.

203. Li, Naicong (1986), "Pronoun Resolution in SNePS," *SNeRG Technical Note No. 18* (Buffalo: SUNY Buffalo Department of Computer Science, SNePS Research Group).

204. Loar, Brian (1982), "Conceptual Role and Truth-Conditions," *Notre Dame Journal of Formal Logic* 23: 272–283.

205. Loewer, Barry (1982), "The Role of 'Conceptual Role Semantics'," *Notre Dame Journal of Formal Logic* 23: 305–315.

206. Lourie, Richard (1992), "Raskolnikov Says the Darndest Things," *New York Times Book Review* (26 April): 24.

207. Lycan, William G. (1984) *Logical Form in Natural Language* (Cambridge, MA: MIT Press).

208. Lyons, John (1973), "Deixis as the Source of Reference," in Edward L. Keenan (ed.), *Formal Semantics of Natural Language* (Cambridge, UK: Cambridge University Press, 1975): 61–83.

209. Lytinen, Steven L. (1992), "Conceptual Dependency and Its Descendants," *Computers and Mathematics with Applications* 23: 51–73.

210. Mac Lane, Saunders (1981), "Mathematical Models: A Sketch for the Philosophy of Mathematics," *American Mathematical Monthly* 88: 462–472.

211. Maida, Anthony S., & Shapiro, Stuart C. (1982), "Intensional Concepts in Propositional Semantic Networks," *Cognitive Science* 6: 291–330; reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 169–189.

212. Malory, Sir Thomas (1470), *Le Morte Darthur*, by R. M. Lumiansky (ed. and trans.) (New York: Collier Books, 1982).

213. Marcotty, Michael, & Ledgard, Henry (1986), *Programming Landscape: Syntax, Semantics, and Implementation*, 2nd edition (Chicago: Science Research Associates).

214. Marsh, B.; Brown, C.; LeBlanc, T.; Scott, M.; Becker, T.; Das, P.; Karlsson, J.; & Quiroz, C. (1992), "Operating System Support for Animate Vision," *Journal of Parallel and Distributed Computing* 15: 103–117.

215. Marsh, Brian; Brown, Chris; LeBlanc, Thomas; Scott, Michael; Becker, Tim; Quiroz, Cesar; Das, Prakash; & Karlsson, Jonas (1992), "The Rochester Checkers Player: Multimodal Parallel Programming for Animate Vision," *Computer*, Vol. ??, No. ?? (February 1992) 12–19.

216. Martins, João, & Shapiro, Stuart C. (1988), "A Model for Belief Revision," *Artificial Intelligence* 35: 25–79.

217. Martins, João, & Cravo, Maria R. (1991), "How to Change Your Mind," *Noûs* 25: 537–551.

218. Mayes, A. R. (1991), Review of Hanna Damasio & Antonio R. Damasio, *Lesion Analysis in Neuropsychology* (*inter alia*), *British Journal of Psychology* 82: 109–112.

219. McCarthy, John (1990), "Syntax, Semantics and Systems," abstract of presentation to the Society for Philosophy and Psychology Symposium on Searle's Chinese Room and Workshop on Symbol Grounding, University of Maryland (electronic mail posting from Stevan Harnad, 6 June 1990).

220. McCloskey 1991

221. Melzack on phantom limbs

222. Miles, H. Lyn White (1990), "The Cognitive Foundations for Reference in a Signing Orangutan," in S. T. Parker & K. R. Gibson (eds.), *"Language" and Intellect in Monkeys and Apes* (Cambridge, UK: Cambridge University Press): 511–539.

223. Meinong, Alexius (1904), "Über Gegenstandstheorie," in Rudolf Haller (ed.), *Alexius Meinong Gesamtausgabe*, Vol. II (Graz, Austria: Akademische Druck- u. Verlagsanstalt, 1971): 481–535; English translation ("The Theory of Objects") by Isaac Levi et al., in Roderick M. Chisholm (ed.), *Realism and the Background of Phenomenology* (New York: Free Press, 1960): 76–117.

224. Minsky, Marvin (1975), "A Framework for Representing Knowledge," in Patrick Henry Winston (ed.), *The Psychology of Computer Vision* (New York: McGraw-Hill); reprinted in John Haugeland (ed.), *Mind Design: Philosophy, Psychology, Artificial Intelligence* (Montgomery, VT: Bradford Books): 95–128; reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 245–262.

225. Minsky, Marvin (1991), posting to `comp.ai.philosophy` bulletin board, 21 December 1991.

226. Moore, G. E. (1939), "Proof of an External World," reprinted in G. E. Moore, *Philosophical Papers* (New York: Collier Books, 1962): 126–148.

227. Moore, Robert C. (1989), "Propositional Attitudes and Russellian Propositions," in R. Bartsch, J. van Benthem, & P. van Emde Boas (eds.), *Semantics and Contextual Expression* (Dordrecht, Holland: Foris Publications): 147–174.

228. Morris, Charles (1938), *Foundations of the Theory of Signs* (Chicago: University of Chicago Press).

229. Neal, Jeannette G. (1981), "A Knowledge Engineering Approach to Natural Language Understanding," *Technical Report 179* (Buffalo: SUNY Buffalo Department of Computer Science).

230. Neal, Jeannette G. (1985), "A Knowledge Based Approach to Natural Language Understanding," *Technical Report 85-06* (Buffalo: SUNY Buffalo Department of Computer Science).

231. Neal, Jeannette G., & Shapiro, Stuart C. (1984), "Knowledge Based Parsing," *Technical Report 213* (Buffalo: SUNY Buffalo Department of Computer Science).

232. Neal, Jeannette G., & Shapiro, Stuart C. (1985), "Parsing as a Form of Inference in a Multiprocessing Environment," *Proceedings of the Conference on Intelligent Systems and Machines (Oakland University, Rochester, MI)*: 19–24.

233. Neal, Jeannette G., & Shapiro, Stuart C. (1987), "Knowledge Based Parsing," in Leonard Bolc (ed.), *Natural Language Parsing Systems* (Berlin: Springer-Verlag).

234. Neal, Jeannette G.; Thielman, C. Y.; Dobes, Zuzanna; Haller, Susan M.; & Shapiro, Stuart C. (1989), "Natural Language with Integrated Deictic and Graphic Gestures," *Proceedings of the DARPA Speech and Natural Language Workshop* (San Mateo, CA: Morgan Kaufmann): 14.

235. Nelson, David A. (1992), "Deductive Program Verification (A Practitioner's Commentary)", *Minds and Machines* 2: 283–307.

236. Nelson, David A. (1994), Review of Robert S. Boyer & J Strother Moore, *A Computational Logic Handbook*, and J Strother Moore, "Special Issue on System Verification", *Minds and Machines* 4: 93–101.

237. Nilsson, Nils J. (1991), "Logic and Artificial Intelligence," *Artificial Intelligence* 47: 31–56.

238. Parnas, David (1972), "A Technique for Software Module Specification with Examples," *Communications of the Association for Computing Machinery* 15: 330–336.

239. Parsons, Terence (1974), "A Prolegomenon to Meinongian Semantics," *Journal of Philosophy* 71: 561–580.

240. Parsons, Terence (1975), "A Meinongian Analysis of Fictional Objects," *Grazer Philosophische Studien* 1: 73–86.

241. Parsons, Terence (1978), "Nuclear and Extranuclear Properties, Meinong and Leibniz," *Noûs* 12: 137–151.

242. Parsons, Terence (1979a), "The Methodology of Nonexistence," *Journal of Philosophy* 76: 649–662.

243. Parsons, Terence (1979b), "Referring to Nonexistent Objects," *Theory and Decision* 11: 95–110.

244. Parsons, Terence (1980), *Nonexistent Objects* (New Haven: Yale University Press).

245. Pavel, Thomas G. (1986), *Fictional Worlds* (Cambridge, MA: Harvard University Press).

246. Perry, John (1979), "The Problem of the Essential Indexical," *Noûs* 13: 3–21.

247. Pelletier, Jeff (1994a), "The Principle of Semantic Compositionality," *Topoi* 13: 11–24.

248. Pelletier, Jeff (1994b), "Semantic Compositionality: The Argument from Synonymy," in R. Casati, B. Smith, & G. White (eds.), *Philosophy and the Cognitive Sciences* (Vienna: Hölder-Pichler-Tempsky): 311–317.

249. Pelletier, Jeff (1994c), "On an Argument against Semantic COmpositionality," in D. Prawitz & D. Westerstahl (eds.), *Logic and Philosophy of Science in Uppsala* (Drodrecht, Holland: Kluwer): 599–610.

250. Peters, Sandra L., & Shapiro, Stuart C. (1987a), "A Representation for Natural Category Systems—I," *Proceedings of the 9th Annual Conference of the Cognitive Science Society (Seattle)* (Hillsdale, NJ: Lawrence Erlbaum Associates): 379–390.

251. Peters, Sandra L., & Shapiro, Stuart C. (1987b), "A Representation for Natural Category Systems—II," *Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI-87, Milan)* (Los Altos, CA: Morgan Kaufmann): 140–146.

252. Peters, Sandra L.; Shapiro, Stuart C.; & Rapaport, William J. (1988), "Flexible Natural Language Processing and Roschian Category Theory," *Proceedings of the 10th Annual Conference of the Cognitive Science Society (Montreal)* (Hillsdale, NJ: Lawrence Erlbaum Associates): 125–131.

253. Peters, Sandra L., & Rapaport, William J. (1990), "Superordinate and Basic Level Categories in Discourse: Memory and Context," *Proceedings of the 12th Annual Conference of the Cognitive Science Society (Cambridge, MA)* (Hillsdale, NJ: Lawrence Erlbaum Associates): 157–165.

254. Perlis, Donald (1991), "Putting One's Foot in One's Head—Part I: Why," *Noûs* 25: 435–455.

255. Perlis, Donald (1994), "Putting One's Foot in One's Head—Part II: How," in Eric Dietrich (ed.), *Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines* (San Diego: Academic Press): 197–224.

256. Pincus, Andrew L. (1990), "The Art of Transcription Sheds New Light on Old Work," *The New York Times*, Arts and Leisure (Sect. 2) (23 September 1990). **pages??**

257. Pinker, Steven, & Mehler, Jacques (eds.) (1988), *Connections and Symbols* (Cambridge, MA: MIT Press).

258. Pollock, Donald (1994), "Personhood and Illness among the Kulina," *Medical Anthropology Quarterly* (forthcoming); page reference to manuscript version.

259. Posner, Roland (1992), "Origins and Development of Contemporary Syntactics," *Languages of Design* 1: 37–50.

260. Potts, Timothy C. (1973), "Model Theory and Linguistics," in Edward L. Keenan (ed.), *Formal Semantics of Natural Language* (Cambridge, UK: Cambridge University Press, 1975): 241–250.

261. Putnam, Hilary (1975), "The Meaning of 'Meaning'," reprinted in *Mind, Language and Reality* (Cambridge, UK: Cambridge University Press): 215–271.

262. Pylyshyn, Zenon (1981), "The Imagery Debate: Analog Media versus Tacit Knowledge," in Ned Block (ed.), *Imagery* (Cambridge, MA: MIT Press): 151–206.

263. Pylyshyn, Zenon (1985), *Computation and Cognition: Toward a Foundation for Cognitive Science, 2nd edition* (Cambridge, MA: MIT Press).

264. Quillian, M. Ross (1967), "Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities," *Behavioral Science* 12: 410–430; reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 97–118.

265. Quillian, M. Ross (1968), "Semantic Memory," in Marvin Minsky (ed.) *Semantic Information Processing* (Cambridge, MA: MIT Press): 227–270.

266. Quine, Willard Van Orman (1951), "Two Dogmas of Empiricism," reprinted in *From a Logical Point of View, 2nd Edition, Revised* (Cambridge, MA: Harvard University Press, 1980): 20–46.

267. Quine, Willard Van Orman, & Ullian, Joseph S. (1978), *The Web of Belief, 2nd Edition* (New York: Random House).

268. Quirke, Stephen, & Andrews, Carol (1988), *The Rosetta Stone: Facsimile Drawing* (New York: Harry N. Abrams, 1989).

269. Rapaport, William J. (1976), *Intentionality and the Structure of Existence*, Ph.D. dissertation (Bloomington: Indiana University Department of Philosophy).

270. Rapaport, William J. (1978), "Meinongian Theories and a Russellian Paradox," *Noûs* 12: 153–180; errata, *Noûs* 13 (1979) 125.

271. Rapaport, William J. (1981), "How to Make the World Fit Our Language: An Essay in Meinongian Semantics," *Grazer Philosophische Studien* 14: 1–21.

272. Rapaport, William J. (1984), Critical Notice of Richard Routley, *Exploring Meinong's Jungle and Beyond*, in *Philosophy and Phenomenological Research* 44: 539–552.

273. Rapaport, William J. (1985a), "To Be and Not to Be: Critical Study of Terence Parsons, *Nonexistent Objects*," *Noûs* 19: 255–271; erratum, *Noûs* 20 (1986) 587.

274. Rapaport, William J. (1985b), "Machine Understanding and Data Abstraction in Searle's Chinese Room," *Proceedings of the 7th Annual Meeting of the Cognitive Science Society (University of California at Irvine)* (Hillsdale, NJ: Lawrence Erlbaum Associates): 341–345.

275. Rapaport, William J. (1985/1986), "Non-Existent Objects and Epistemological Ontology," *Grazer Philosophische Studien* 25/26: 61–95; reprinted in Rudolf Haller (ed.), *Non-Existence and Predication* (Amsterdam: Rodopi, 1986).

276. Rapaport, William J. (1986a), "Logical Foundations for Belief Representation," *Cognitive Science* 10: 371–422.

277. Rapaport, William J. (1986b), "Philosophy, Artificial Intelligence, and the Chinese-Room Argument," *Abacus* 3 (Summer 1986) 6–17; correspondence, *Abacus* 4 (Winter 1987) 6–7, *Abacus* 4 (Spring 1987) 5–7.

278. Rapaport, William J. (1986c), "Searle's Experiments with Thought," *Philosophy of Science* 53: 271–279; preprinted as *Technical Report 216* (Buffalo: SUNY Buffalo Department of Computer Science, 1984).

279. Rapaport, William J. (1988a), "To Think or Not to Think," *Noûs* 22: 585–609.

280. Rapaport, William J. (1988b), "Syntactic Semantics: Foundations of Computational Natural-Language Understanding," in James H. Fetzer (ed.), *Aspects of Artificial Intelligence* (Dordrecht, Holland: Kluwer Academic Publishers): 81–131; reprinted in Eric Dietrich (ed.), *Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines* (San Diego: Academic Press, 1994): 225–273.

281. Rapaport, William J. (1990), "Computer Processes and Virtual Persons: Comments on Cole's 'Artificial Intelligence and Personal Identity'," *Technical Report 90-13* (Buffalo: SUNY Buffalo Department of Computer Science, May 1990).

282. Rapaport, William J. (1991a), "Predication, Fiction, and Artificial Intelligence," *Topoi* 10: 79–111; pre-printed as *Technical Report 90-11* (Buffalo: SUNY Buffalo Department of Computer Science, May 1990).

283. Rapaport, William J. (1991b), "Meinong, Alexius I: Meinongian Semantics," in Hans Burkhardt & Barry Smith (eds.), *Handbook of Metaphysics and Ontology* (Munich: Philosophia Verlag): 516–519.

284. Rapaport, William J. (1992a), "Logic, Propositional," in Stuart C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence, 2nd edition* (New York: John Wiley): 891–897.

285. Rapaport, William J. (1992), "Logic, Predicate," in Stuart C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence, 2nd edition* (New York: John Wiley): 866–873.

286. Rapaport, William J. (1993a), "Cognitive Science," in Anthony Ralston & Edwin D. Reilly (eds.), *Encyclopedia of Computer Science, 3rd Edition*, (New York: Van Nostrand Reinhold): 185–189.

287. Rapaport, William J. (1993b), "Because Mere Calculating Isn't Thinking: Comments on Hauser's 'Why Isn't My Pocket Calculator a Thinking Thing?'," *Minds and Machines* 3: 11–20.

288. Rapaport, William J.; Segal, Erwin M.; Shapiro, Stuart C.; Zubin, David A.; Bruder, Gail A.; Duchan, Judith F.; Almeida, Michael J.; Daniels, Joyce H.; Galbraith, Mary M.; Wiebe, Janyce M.; & Yuhan, Albert Hanyong (1989), "Deictic Centers and the Cognitive Structure of Narrative Comprehension," *Technical Report 89-01* (Buffalo: SUNY Buffalo Department of Computer Science).

289. Rapaport, William J.; Segal, Erwin M.; Shapiro, Stuart C.; Zubin, David A.; Bruder, Gail A.; Duchan, Judith F.; & Mark, David M. (1989), "Cognitive and Computer Systems for Understanding Narrative Text," *Technical Report 89-07* (Buffalo: SUNY Buffalo Department of Computer Science).

290. Rapaport, William J., & Shapiro, Stuart C. (1984), "Quasi–Indexical Reference in Propositional Semantic Networks," *Proceedings of the 10th International Conference on Computational Linguistics (COLING–84, Stanford University)* (Morristown, NJ: Association for Computational Linguistics): 65–70.

291. Rapaport, William J., & Shapiro, Stuart C. (1995), "Cognition and Fiction," in Judith F. Duchan, Gail A. Bruder, & Lynne E. Hewitt (eds.), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates).

292. Rapaport, William J.; Shapiro, Stuart C.; & Wiebe, Janyce M. (forthcoming), "Quasi-Indexicals and Knowledge Reports," *Cognitive Science*: 000–000.

293. Récanati, François (1995), "The Alleged Priority of Literal Interpretation," *Cognitive Science* 19: 207–232.

294. Renfrew, Colin (1990), "Towards a Cognitive Archeology," in "For Participants in the Workshop for an Explicitly Scientific Cognitive Archeology," unpublished ms. (Cambridge, UK: University of Cambridge Department of Archeology, 8–13 April 1990).

295. Roberts, Lawrence D., & Rapaport, William J. (1988), "Quantifier Order, Reflexive Pronouns, and Quasi-Indexicals," *Technical Report 88-16* (Buffalo: SUNY Buffalo Department of Computer Science, August 1988).

296. Rosch, Eleanor (1978), "Principles of Categorization," in Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and Categorization* (Hillsdale, NJ: Lawrence Erlbaum): 27–48.

297. Rosen, Charles (1991), Reply to letter, *New York Review of Books* (14 February 1991): 50.

298. Rosenblueth, Arturo, & Wiener, Norbert (1945), "The Role of Models in Science," *Philosophy of Science* 12: 316–321.

299. Routley, Richard (1979), *Exploring Meinong's Jungle and Beyond* (Canberra: Australian National University, Research School of Social Sciences, Department of Philosophy).

300. Russell, Bertrand (1901), "Mathematics and the Metaphysicians," reprinted in Bertrand Russell (1917), *Mysticism and Logic* (New York: Doubleday Anchor Books, 1957): 70–92.

301. Russell, Bertrand (1902), "The Study of Mathematics," first published in 1907; reprinted in Bertrand Russell (1917), *Mysticism and Logic* (New York: Doubleday Anchor Books, 1957): 55–69.

302. Russell, Bertrand (1903), *The Principles of Mathematics* (New York: W. W. Norton, 19??).

303. Russell, Bertrand (1905), "On Denoting," reprinted in *Logic and Knowledge*, R. C. Marsh (ed.), (New York: Capricorn, 1956): 39–56.

304. Russell, Bertrand (1918), logical atomism? reprinted in *Logic and Knowledge*, R. C. Marsh (ed.), (New York: Capricorn, 1956): 000–000.

305. Sacks, Oliver W. (1989), *Seeing Voices: A Journey into the World of the Deaf* (Berkeley: University of California Press).

306. Sacks, Oliver (1990), "Seeing Voices," *Exploratorium Quarterly* Vol. 14, No. 2 (Summer), p. 3.

307. Sagan, Carl (1980), *Cosmos* (New York: Random House).

308. Sayre, Kenneth M. (1973), "Machine Recognition of Handwritten Words," *Pattern Recognition Journal* 5: 213–228.

309. Schank, Roger C. (1975), *Conceptual Information Processing* (New York: Elsevier).

310. Schank, Roger C., & Rieger, Charles J. (1974), "Inference and the Computer Understanding of Natural Language," *Artificial Intelligence* 5: 373–412.

311. Schank, Roger C., & Riesbeck, Christopher K. (eds.) (1981), *Inside Computer Understanding: Five Programs Plus Miniatures* (Hillsdale, NJ: Lawrence Erlbaum).

312. Schiller, F. C. S. (anonymously) (1901), *Mind! A Unique Review of Ancient and Modern Philosophy. Edited by A Troglodyte, With the Co-operation of the Absolute and others* (London: Williams & Norgate).

313. Schneiderman, Ben (1993), "Data Type," in Anthony Ralston & Edwin D. Reilly (eds.), *Encyclopedia of Computer Science, 3rd edition*, (New York: Van Nostrand Reinhold): 411–412.

314. Schonberg, Harold C. (1990), "Some Chessmen Don't Make a Move," *New York Times* (15 April 1990), Sect. 2, pp. 38–39.

315. Schulberg, Budd (1995), "First a Movie, Then a Novel, Now a Play," *New York Times* (30 April), Sect. H, pp. 5, 10.

316. Searle, John R. (1979), "The Logical Status of Fictional Discourse," in *Expression and Meaning* (Cambridge, UK: Cambridge Univesity Press): 58–75.

317. Searle, John R. (1980), "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3: 417–457.

318. Searle, John R. (1982), "The Myth of the Computer," *New York Review of Books* (29 April 1982): 3–6; cf. correspondence, same journal (24 June 1982): 56–57.

319. Searle, John R. (1984), *Minds, Brains and Science* (Cambridge, MA: Harvard University Press).

320. Searle, John R. (1990), "Is the Brain a Digital Computer?", *Proceedings and Addresses of the American Philosophical Association*, Vol. 64, No. 3: 21–37.

321. Searle, John R. (1993), "The Failures of Computationalism," *Think* (Tilburg, The Netherlands: Tilburg University Institute for Language Technology and Artificial Intelligence) 2 (June 1993) 68–71.

322. Segal, Erwin M.; Duchan, Judith F.; & Scott, Paula J. (1991), "The Role of Interclausal Connectives in Narrative Structuring: Evidence from Adults' Interpretations of Simple Stories," *Discourse Processes* 14: 27–54.

323. Segal, Erwin M. (1995), "Stories, Story Worlds, and Narrative Discourse," in Judith F. Duchan, Gail A. Bruder, & Lynne E. Hewitt (eds.), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates).

324. Sellars, Wilfrid (1955/1963), "Some Reflections on Language Games," in *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963): 321–358.

325. Sellars, Wilfrid (1959/1963), "The Language of Theories," in *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963): 106–126.

326. Sellars, Wilfrid (1961/1963), "Truth and 'Correspondence'," in *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963): 197–224.

327. Shannon, Claude E. (1949), *The Mathematical Theory of Communication* (Urbana: University of Illinois Press, 1964).

328. Shapiro, Stuart C. (1977), "Representing Numbers in Semantic Networks: Prolegomena," *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI-77, MIT)* (Los Altos, CA: Morgan Kaufmann): 284.

329. Shapiro, Stuart C. (1979), "The SNePS Semantic Network Processing System," in Nicholas Findler (ed.), *Associative Networks: Representation and Use of Knowledge by Computers* (New York: Academic Press): 179–203.

330. Shapiro, Stuart C. (1982), "Generalized Augmented Transition Network Grammars for Generation from Semantic Networks," *American Journal of Computational Linguistics* 8: 12–25.

331. Shapiro, Stuart C. (1986), "Symmetric Relations, Intensional Individuals, and Variable Binding," *Proceedings of the IEEE* 74: 1354–1363.

332. Shapiro, Stuart C. (1989), "The CASSIE Projects: An Approach to Natural Language Competence," *Proceedings of the 4th Portugese Conference on Artificial Intelligence (Lisbon)* (CITY: Springer-Verlag): 362–380.

333. Shapiro, Stuart C. (1992), "Artificial Intelligence," in Stuart C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence, 2nd Edition* (New York: John Wiley & Sons): 54–57.

334. Shapiro, Stuart C. (1993), "Belief Spaces as Sets of Propositions," *Journal of Experimental and Theoretical Artificial Intelligence* 5: 225–235.

335. Shapiro, Stuart C., & Rapaport, William J. (1987), "SNePS Considered as a Fully Intensional Propositional Semantic Network," in Nick Cercone & Gordon McCalla (eds.), *The Knowledge Frontier: Essays in the Representation of Knowledge* (New York: Springer-Verlag): 262–315; earlier version preprinted as *Technical Report 85-15* (Buffalo: SUNY Buffalo Department of Computer Science, 1985); shorter version appeared in *Proceedings of the 5th National Conference on Artificial Intelligence (AAAI-86, Philadelphia)* (Los Altos, CA: Morgan Kaufmann): 278–283; revised version of the shorter version appears as "A Fully Intensional Propositional Semantic Network," in Leslie Burkholder (ed.), *Philosophy and the Computer* (Boulder, CO: Westview Press, 1992): 75–91.

336. Shapiro, Stuart C., & Rapaport, William J. (1991), "Models and Minds: Knowledge Representation for Natural-Language Competence," in Robert Cummins & John Pollock (eds.), *Philosophy and AI: Essays at the Interface* (Cambridge, MA: MIT Press): 215–259.

337. Shapiro, Stuart C., & Rapaport, William J. (1992), "The SNePS Family," *Computers and Mathematics with Applications* 23: 243–275; reprinted in Fritz Lehmann (ed.), *Semantic Networks in Artificial Intelligence* (Oxford: Pergamon Press, 1992): 243–275.

338. Shapiro, Stuart C., & Rapaport, William J. (1995), "An Introduction to a Computational Reader of Narrative," in Judith F. Duchan, Gail A. Bruder, & Lynne E. Hewitt (eds.), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates): 79–105.

339. Sheckley, Robert (1954), "Ritual," in *Untouched by Human Hands* (New York: Ballantine Books): 155–165.

340. Shoemaker on qualia

341. Sidner, Candace L. (1994), "An Artificial Discourse Language for Collaborative Negotiation," *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94, Seattle)* (Menlo Park, CA: AAAI Press/MIT Press): 814–819.

342. Simon, Herbert A. (1977), "Artificial Intelligence Systems that Understand," *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI-77, MIT)* (Pittsburgh, PA: Carnegie-Mellon University Department of Computer Science): 1059–1073.

343. Simon, Herbert A. (1992), "The Computer as a Laboratory for Epistemology," in Leslie Burkholder (ed.), *Philosophy and the Computer* (Boulder, CO: Westview Press): 3–23.

344. Simpson, J. A., & Weiner, E. S. C. (preparers) (1989), *The Oxford English Dictionary, 2nd edition* (Oxford: Clarendon Press).

345. Sloman, Aaron (1985), "What Enables a Machine to Understand?", *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI-85, UCLA)* (Los Altos, CA: Morgan Kaufmann): 995–1001.

346. Smith, Brian C. (1982), "Reflection and Semantics in a Procedural Language," *Technical Report MIT/LCS/TR-272* (Cambridge, MA: MIT Laboratory for Computer Science); page

reference is to the prologue, reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 32–39.

347. Smith 15?

348. ? Smith, Brian Cantwell (1982a/104), "Semantic Attribution and the Formality Condition," unpublished draft of paper presented at the 8th Annual Meeting of the Society for Philosophy and Psychology, University of Western Ontario.

349. Smith, Brian Cantwell (1982b), "Linguistic and Computational Semantics," *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics (University of Toronto)* (Morristown, NJ: Association for Computational Linguistics): 9–15.

350. Smith, Brian Cantwell (1985), "Limits of Correctness in Computers," in Charles Dunlop & Rob Kling (eds.), *Computerization and Controversy* (San Diego: Academic Press, 1991): 632–646.

351. Smith, Brian Cantwell (1987), "The Correspondence Continuum," *Report CSLI-87-71* (Stanford, CA: Center for the Study of Language and Information).

352. Smith, Brian Cantwell (1991), "The Owl and the Electric Encyclopedia," *Artificial Intelligence* 47: 251–288.

353. Smolensky, Paul (1988), "The Proper Treatment of Connectionism," *Behavioral and Brain Sciences* 11: 1–74.

354. Sowa, John F. (1984), *Conceptual Structures: Information Processing in Mind and Machine* (Reading, MA: Addison-Wesley).

355. Sowa, John F. (1992), "Conceptual Graphs as a Universal Knowledge Representation," *Computers and Mathematics with Applications* 23: 75–93.

356. Srihari, Rohini K. (1991a), "PICTION: A System that Uses Captions to Label Human Faces in Newspaper Photographs," *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI-91, Anaheim)* (Menlo Park, CA: AAAI Press/MIT Press): 80–85.

357. Srihari, Rohini K. (1991b), "Extracting Visual Information from Text: Using Captions to Label Faces in Newspaper Photographs," *Technical Report 91-17 (Buffalo: SUNY Buffalo Department of Computer Science*.

358. Srihari, Rohini K. (1993a), "Intelligent Document Understanding: Understanding Photos with Captions," *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR-93, Tsukuba City, Japan)*, forthcoming.

359. Srihari, Rohini K. (1993b), "Use of Collateral Text in Understanding Photos in Documents," *Proceedings of the Conference on Applied Imagery and Pattern Recognition (AIPR/SPIE, Washington, DC)*, forthcoming.

360. Srihari, Rohini K., & Rapaport, William J. (1989), "Extracting Visual Information From Text: Using Captions to Label Human Faces in Newspaper Photographs," *Proceedings of the 11th Annual Conference of the Cognitive Science Society (Ann Arbor, MI)* (Hillsdale, NJ: Lawrence Erlbaum Associates): 364–371.

361. Srihari, Rohini K., & Rapaport, William J. (1990), "Combining Linguistic and Pictorial Information: Using Captions to Interpret Newspaper Photographs," in Deepak Kumar (ed.), *Current Trends in SNePS—Semantic Network Processing System*, Lecture Notes in Artificial Intelligence, No. 437 (Berlin: Springer-Verlag): 85–96.

362. Steenrod, N. (1967), "The Geometric Content of Freshman and Sophomore Mathematics Courses," *CUPM Geometry Conference Report II* (October 1967): 1–52; cited in Murray S. Klamkin, Review of A. Soifer, *How Does One Cut a Triangle?*, in *American Mathematical Monthly* 98 (1991) 775–778.

363. Stich, Stephen P. (1978), "Autonomous Psychology and the Belief–Desire Thesis," *The Monist* 61: 573–591.

364. Stich, Stephen (1983), *From Folk Psychology to Cognitive Science: The Case Against Belief* (Cambridge, MA: MIT Press).

365. Suits, David B. (1989), "Out of the Chinese Room," *The Computers and Philosophy Newsletter*, Issue 4:1+4:2 (July), pp. 1–7.

366. Swan, Jim (1994), "Touching Words: Helen Keller, Plagiarism, Authorship," in Martha Woodmansee & Peter Jaszi (eds.), *The Construction of Authorship: Textual Appropriation in Law and Literature* (Durham, NC: Duke University Press): 57–100.

367. Talmy, Leonard (1978), "Figure and Ground in Complex Sentences," in Joseph H. Greenberg (ed.), *Universals of Human Language*, Vol. 4: *Syntax* (Stanford, CA: Stanford University Press): 625–649.

368. Tanenbaum, Andrew S. (1976), *Structured Computer Organization* (Englewood Cliffs, NJ: Prentice-Hall).

369. Tenenbaum, Aaron M., & Augenstein, Moshe J. (1981), *Data Structures using Pascal* (Englewood Cliffs, NJ: Prentice-Hall).

370. Terrace, Herbert S. (1985), "In the Beginning Was the 'Name'," *American Psychologist* 40: 1011–1028.

371. Terrace, Herbert S. (1991), Letter to the Editor, *New York Review of Books*, Vol. 38, No. 15 (10 October 1991) 53.

372. Turner, Michael (1987), *Death Is the Mother of Beauty* (Chicago: University of Chicago Press).

373. Twardowski, Kasimir (1894), *On the Content and Object of Presentations*, Reinhardt Grossmann (trans.) (The Hague: Nijhoff, 1977).

374. Vauclair, Jacques (1990), "Primate Cognition: From Representation to Language," in S. T. Parker & K. R. Gibson (eds.), *"Language" and Intellect in Monkeys and Apes* (Cambridge, UK: Cambridge University Press): 312–329.

375. Von Glasersfeld, E. (1977), "Linguistic Communication: Theory and Definition," in Duane M. Rumbaugh (ed.), *Language Learning by a Chimpanzee: The LANA Project* (New York: Academic Press): 55–71.

376. Wartofsky, Marx W. (1966), "The Model Muddle: Proposals for an Immodest Realism," in *Models: Representation and the Scientific Understanding* (Dordrecht, Holland: D. Reidel, 1979): 1–11.

377. Wartofsky, Marx W. (1979), "Introduction," in *Models: Representation and the Scientific Understanding* (Dordrecht, Holland: D. Reidel, 1979): xiii—xxvi.

378. Webb, B. H. (1991), "Do Computer Simulations Really Cognize?", *Journal of Experimental and Theoretical Artificial Ingelligence* 3: 247–254.

379. Weizenbaum, Joseph (1976), *Computer Power and Human Reason: From Judgment to Calculation* (New York: W. H. Freeman).

380. Wiebe, Janyce M. (1990), "Recognizing Subjective Sentences: A Computational Investigation of Narrative Text," *Technical Report 90-03* (Buffalo: SUNY Buffalo Department of Computer Science).

381. Wiebe, Janyce M. (1991), "References in Narrative Text", *Noûs*, 25: 457–486; reprinted in Judith F. Duchan, Gail A. Bruder, & Lynne E. Hewitt (eds.), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1995): 263–286.

382. Wiebe, Janyce M. (1994), "Tracking Point of View in Narrative," *Computational Linguistics* 20: 233–287.

383. Wiebe, Janyce M., & Rapaport, William J. (1986), "Representing *De Re* and *De Dicto* Belief Reports in Discourse and Narrative," *Proceedings of the IEEE* 74: 1405–1413.

384. Wiebe, Janyce M., & Rapaport, William J. (1988), "A Computational Theory of Perspective and Reference in Narrative" *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (SUNY Buffalo)* (Morristown, NJ. Association for Computational Linguistics): 131–138.

385. Wilensky, Robert (1991), "Sentences, Situations, and Propositions," in John F. Sowa (ed.), *Principles of Semantic Networks: Explorations in the Representation of Knowledge* (San Mateo, CA: Morgan Kaufmann): 191–227.

386. Wilkins, David P. (1991), "The Semantics, Pragmatics and Diachronic Development of 'Associated Motion' in Mparntwe Arrernte," in Dan Devitt, Fengxiang Li, Heater L. Weber, Robert D. Van Valin Jr., & Lindsay J. Whaley (eds.), *Buffalo Papers in Linguistics BPL 91-01* (Buffalo: SUNY Buffalo Department of Linguistics): 207–257.

387. Wilkins, David P. (1992), "Interjections as Deictics," *Journal of Pragmatics* 18:"119–158.

388. Wilkins, David P. (1995), "Expanding the Traditional Category of Deictic Elements: Interjections as Deictics," in Judith F. Duchan, Gail A. Bruder, & Lynne E. Hewitt (eds.), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1995): 359–386.

389. Wilks, Yorick (1971), "Decidability and Natural Language," *Mind* 80: 497–520.

390. Wilks, Yorick (1972), *Grammar, Meaning and the Machine Analysis of Language* (London: Routledge & Kegan Paul).

391. Wilks, Yorick (1975), "An Intelligent Analyzer and Understander of English," *Communications of the Association for Computing Machinery* 18: 264–274; reprinted in Barbara J. Grosz, Karen Sparck Jones, & Bonnie L. Webber (eds.), *Readings in Natural Language Processing* (Los Altos, CA: Morgan Kaufmann, 1986): 193–203.

392. Wilks, Yorick, & Fass, Dan (1992), "The Preference Semantics Family," *Computers and Mathematics with Applications* 23: 205–221; reprinted in Fritz Lehmann (ed.), *Semantic Networks in Artificial Intelligence* (Oxford: Pergamon Press, 1992): 205–221.

393. Willoughby, Stephen S. (1967), *Contemporary Teaching of Secondary School Mathematics* (New York: John Wiley & Sons).

394. Wills, Gary (1991), "Man of the Year," *New York Review of Books* (21 November 1991): 12, 14–18.

395. Winograd, Terry (1972), *Understanding Natural Language* (Orlando, FL: Academic Press).

396. Winograd, Terry (1975), "Frame Representations and the Declarative/Procedural Controversy," in Daniel G. Bobrow & Alan M. Collins (eds.), *Representation and Understanding: Studies in Cognitive Science* (New York: Academic Press): 185–210; reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 358–370.

397. Winston, Patrick Henry (1975), "Learning Structural Descriptions from Examples," in Patrick Henry Winston (ed.), *The Psychology of Computer Vision* (New York: McGraw-Hill): 157–209; reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 141–168 (page references are to this reprint).

398. Winston, Patrick Henry (1977), *Artificial Intelligence* (Reading, MA: Addison-Wesley).

399. Wittgenstein, Ludwig (1958), *Philosophical Investigations: The English Text of the Third Edition*, trans. by G. E. M. Anscombe (New York: Macmillan).

400. Wolff, J. G. (1978), "Grammar Discovery as Data Compression," *Proceedings of the AISB/GI Conference on Artificial Intelligence (Hamburg, W. Germany)*: 375–379.

401. Wolff, J. G. (1982), "Language Acquisition, Data Compression, and Generalization," *Language and Communication* 2: 57–89.

402. Wolterstorff, Nicholas (1970), "Bergmann's Constituent Ontology," *Noûs* 4: 109–134.

403. Woods, William A. (1975), "What's in a Link: Foundations for Semantic Networks," in Daniel G. Bobrow & Alan M. Collins (eds.), *Representation and Understanding: Studies in Cognitive Science* (New York: Academic Press): 35–82; reprinted in Ronald J. Brachman & Hector J. Levesque (eds.), *Readings in Knowledge Representation* (Los Altos, CA: Morgan Kaufmann, 1985): 217–241.

404. Woods, William A., & Schmolze, James G. (1992), "The KL-ONE Family," *Computers and Mathematics with Applications* 23: 133–177; reprinted in Fritz Lehmann (ed.), *Semantic Networks in Artificial Intelligence* (Oxford: Pergamon Press, 1992): 133–177.

405. Wyatt, Richard (1989), "The Representation of Opaque Contexts," *Technical Report 89-13* (Buffalo: SUNY Buffalo Department of Computer Science).

406. Wyatt, Richard (1990), "Kinds of Opacity and Their Representations," in Deepak Kumar (ed.), *Current Trends in SNePS—Semantic Network Processing System*, Lecture Notes in Artificial Intelligence, No. 437 (Berlin: Springer-Verlag): 123–144.

407. Wyatt, Richard (1993), "Reference and Intensions," *Journal of Experimental and Theoretical Artificial Intelligence* 5: 263–271.

408. Yuhan, Albert Hanyong (1991), "Dynamic Computation of Spatial Reference Frames in Narrative Understanding," *Technical Report 91-03* (Buffalo: SUNY Buffalo Department of Computer Science).

409. Yuhan, Albert Hanyong, & Shapiro, Stuart C. (1995), "Computational Representation of Space," in Judith Felson Duchan, Gail A. Bruder, & Lynne E. Hewitt (eds.), *Deixis in Narrative: A Cognitive Science Perspective* (Hillsdale, NJ: Lawrence Erlbaum Associates): 191–225.

410. Zadrozny, Wlodek (1994), "From Compositional to Systematic Semantics," *Linguistics and Philosophy* 17: 329–342.

411. Zadrozny, Wlodek, & Jensen, Karen (1991/133), "Semantics of Paragraphs," *Computational Linguistics* 17: 171–209.

412. Zalta, Edward N. (1983), *Abstract Objects: An Introduction to Axiomatic Metaphysics* (Dordrecht, Holland: D. Reidel).

413. Zuckermann, Lord (1991), Letter to the Editor, *New York Review of Books*, Vol. 38, No. 15 (10 October 1991) 53.