1 Platonic Heaven

Sometimes you are drawn to something by pure mystification. The biologist might be baffled by the emergence of life from brute matter. How could it be that a cell could support life? The neuroscientist might be dazzled by the emergence of thought and consciousness from the neurochemistry and topology of the nervous system. How is it that light, hitting the eye, results in the *experience* of red?

The world is surely filled with more than enough to dumbfound and amaze for a lifetime. But the object of wonder that has most enticed me is meaning. What is meaning, and how is it possible for me to mean something? Particularly puzzling is the fact that I can mean something by making noises with my mouth, or making marks on paper, or moving my hands in particular ways. In this chapter, I lay out a way of thinking about meaning in language that has motivated an enormous amount of work in theoretical linguistics. I used to believe fervently in it, but I don't anymore. For reasons I give in chapter 2, I've become a sceptic. In chapter 3, I motivate an economic theory of meaning that lays the foundations for the work I really want to talk about: using the theory of games to think about strategic aspects of meaning.

The Puzzle of Reference

For the moment, though, let's revel in the mystification of meaning. Once, I had to go to a conference in Prague, a city I had heard *spoken of*, had *read about*, but had never seen. I told a travel agent that I had to go to Prague, and after some more noises, I found myself in possession of an airplane ticket. I made a phone call and spoke to some unseen and, to me at least, unknown person who purported to work in a hotel in Prague. I was told, after some negotiating, that I would have a room. All this by moving my tongue and jaw appropriately.

Of course, I wasn't out of the woods yet; perhaps I had accidently arranged to go to Cleveland. I went to the airport, and there was a plane putatively destined for Prague. Apparently, my discussion with the travel agent had worked; the next thing I knew, I was on the plane in exactly the seat that my travel agent told me she had reserved in my name. The plane took off, and there I was with the presumptive destination of Prague, capital of the Czech Republic.

Once I had landed and cleared customs, I found a cab and *told* the cab driver the address of my hotel. And the honest fellow drove me right to it. There it was exactly as promised. Astonishing! Not only that, but I did indeed have a room there, just as I had negotiated with the clerk by telephone.

Later, with the aid of a map and a guidebook, I confirmed that I was in Prague. There was a river precisely where the map promised the Vltava River would be. I walked across a bridge that purported to be the Charles Bridge and saw Prague Castle up on top of a hill, just where it was supposed to be. Using material in the packet the conference organizers had sent me, I went to an address and found a room where a group of people led me to believe that they were attending precisely the same conference I was supposed to be attending. At the appointed time, I gave a talk. The audience nodded, seeming to understand me. Some even asked questions that were relevant to what I had said; apparently I had communicated something to them. They gave every appearance of grasping my meaning.

There were only two possibilities. One possibility was that somehow, using marks and noises, I had successfully gone to Prague, not Cleveland, and given my talk at the correct conference. A more sinister possibility was that I had fallen victim to an immense conspiracy, some vast prank, and had wandered into some hoax Prague perhaps Cleveland disguised as Prague where people pretended to be attending the conference and feigned that they were following my talk.

Rejecting the second possibility as too fantastic who would benefit from such a conspiracy? I settled on the reasonable hypothesis that I was in Prague, the capital of the Czech Republic, attending the conference. I had spoken; my meaning had been understood. Somehow I had used noises to solve problems. How can it be?

Use, Mention, and Truth

The reader may well be baffled by my bafflement, but bear with me. There are good autobiographical reasons why I am so puzzled by simple things.

I was born in the southwest of the United States, near the border with Mexico, at a time when *gringos* mainly white, English-speaking Yankees were a minority. Most people living there were *chicanos*, although we referred to them as *Mexicans* because they were Hispanic and many of them spoke Spanish. Of course, they had been living in the area since Methuselah was in short pants. We *gringos* were the interlopers, and we were decidedly the linguistic minority; English was a relatively recent transplant to the area, Spanish having been imported there centuries earlier. In reality, my family was living in a colonial situation. The white, English-speaking minority community was economically dominant and largely insulated from the poorer, Hispanic majority. Although my parents were far from wealthy, we could easily afford to have a maid come in from Juárez to do the housekeeping.[1]

A few *chicanos* were able to climb into the middle class; I don't remember my parents socializing with them beyond the requisite low-level civilities. I, however, happily played with the Gonzalez children next door, so much so that my parents, already worried by the pervasiveness of Spanish, grew concerned that I might end up speaking better Spanish than English. They needn't have worried. Everything about the social environment and the local economy at the time pushed me toward learning English. English, after all, was the language of the economically and politically dominant class. For years, I associated Spanish with poverty, the language of the underclass.

As a boy, I was surrounded, or so it seemed to me, by the largely impenetrable code of Spanish. When I was small, my mother and I would venture out of our little Anglophonic island and suddenly be immersed in a completely mysterious world where everyone spoke a language we didn't understand. When I was older and could venture out on my own, I learned to curse in Spanish, but was otherwise oblivious to it.

Of course, we had obligatory Spanish lessons throughout primary school. I managed to learn very little, but I did take special note of facts like the following:

(1) *Perro* means dog.

As it happened, I completely misunderstood the translation rules like the one in (1). Instead, I understood them as

(2) *Perro* means *dog.*

There's a crucial difference between (1) and (2), one that had a big impact on me.

Philosophers and linguists make a distinction between *using* a word and *mentioning* it. In (3) I use the word *dog* to help me refer to some particular dog:

(3) My neighbor's dog barked all night.

In (4), I use *dog* to refer to the word itself:

(4) *Dog* is a word of one syllable.

That is, I mention the word *dog.* Clearly, it makes no sense to suppose that actual dogs are monosyllabic words. I've actually mentioned *dog* several times on this page; the mention of a word shows the curious ability of language to turn on itself and talk (and think) about itself.

This ability of language to refer to itself has been a source of enormous philosophical puzzlement, as (5) shows:

(5) This sentence is false.

The sentence in (5) is true if it's false and false if it's true. This is an example of the famous liar paradox, which is often taken to be a problem of self-reference, that is, the word *this* in (5) refers to the sentence that contains it. But really the problem lies in language's ability to talk about language, so *self-reference* must be taken in a very broad sense. To see this, look at the sentences in (6):

(6) a. The sentence in (6b) is true.
    b. The sentence in (6a) is false.

Neither sentence in (6) refers to itself, but they still have the flavor of the liar paradox in (5). If (6a) is true, then it must be false. (6a) asserts that (6b) is true. But (6b) says that (6a) is false. So if (6a) is true, then (6b) must be true and (6a) must be false. The two sentence consume each other like an ouroboros. The problem is that we're using language to talk about itself.

All this is just to say that my early confusion has a distinguished philosophical pedigree. Let's take a closer look at my problem. I understood the teacher as saying

(7) *Perro* means *dog.*

This means not that *perro* means in Spanish what *dog* means in English (an actual canine). Instead, it means that the Spanish word *perro* denotes the English word *dog.*

The Language of Thought

Laugh, if you will, at my boyhood theory of Spanish, but I note that it is not without precedent, and in many ways it is an instance of a perfectly respectable theory of meaning. What I decided was that speakers of Spanish were internally speaking English and translating from English to Spanish when they spoke and from Spanish to English when they listened.

Of course, I had to solve the problem of why Spanish speakers often didn't understand English. I concluded that although they were thinking in English, these English thought processes were inaccessible to their conscious minds (I must have absorbed some talk of Freud and the unconscious). So here was my theory. Although Spanish speakers thought in English, English was not accessible to them as a means of communication. They therefore had to frame everything in terms of the language they knew, namely, Spanish. The grammar of Spanish, then, must be a translation manual between Spanish and English.

Imagine how gratified I was to learn, many years later, that a famous Enlightenment philosopher had asserted that French was the language of thought. When asked whether the ancient Romans thought in French, he unflinchingly responded that they must have done so, even though French is a descendant of Latin. I admire his confidence.

It might seem peculiar to say that we arrive at a speaker's meaning by translating what she says into some other language. But if you think about it, if you can translate correctly from, say, Spanish to English or French to Latin, you would need a very thorough understanding of Spanish or French and English or Latin. Furthermore, if I happen to speak English or Latin, then your translation is very informative. The idea of producing a *translation manual*, a manual for translating from a language you don't know to a language you do know, as a kind of theory of meaning has been advocated by the philosopher W.V.O. Quine, for example. In one form or another, the idea has been a very important one for linguists working on meaning, although I think that Quine would probably disagree with the form much of this work has taken.

Now, where I went wrong, some would say, is not in the idea that speakers of Spanish were translating back and forth between some internal mental language and Spanish, but in supposing that English speakers weren't doing so. What if they were translating from their external language, a language that they would have to learn, into a special internal

language, a language that they understood from birth and thus didn't need to learn. It might be that there is a kind of internal mental language — a Language of Thought (LOT), or Mentalese, as it is sometimes known — and that when people speak they take a sentence in Mentalese and translate it into whatever language they use to communicate. When they hear a sentence in their external language, they analyze it and translate it into Mentalese.

Of course, no one would necessarily have any direct perception of Mentalese. I have a strong intuition that I think in English, but perhaps I'm aware of my thoughts only after they've been translated from Mentalese to English. All sorts of things go on below the level of conscious awareness. For example, I have no reliable intuitions about how I process visual information. I was surprised to learn that the brain has two different visual systems; one system recognizes where objects are in space, and the other system recognizes what the objects are. The two systems can be independently impaired. Someone with an impairment in the "where system" can recognize an object but can't reliably reach for it; someone with an impairment in the "what system" can reach for the object accurately but can't recognize what it is, even though he may know a lot about the kind of thing the object is.

My point is that brute intuitions about plausibility are not the most reliable way to judge an idea. Instead, we need to think about the empirical consequences of the idea. If the theory fails empirically, then we need to cast about for a better theory. Equally, if there's no way to test the theory — no evidence that could possibly count against the theory — then the theory needs to be rejected. Linguists are in the business of producing theories that can be tested empirically.

So think about the following idea. In understanding a sentence, we translate that sentence into the Language of Thought, and when we want to communicate an idea we translate from the Language of Thought into whatever spoken language we use. Assume that part of our linguistic ability includes rules like the following, a *truth predicate*:

(8) S is true $\Leftrightarrow$ Σ.

The *S* in (8) would be a sentence in some natural language like English and the Σ would be a sentence in Mentalese. The double arrow $\Leftrightarrow$ means 'a systematic mapping between', so when I encounter the sentence *S*, I can apply the procedure indicated by $\Leftrightarrow$ and get the resulting expression Σ in Mentalese.

The basic idea is that the way to work out a theory of meaning for a language like English is to show how to translate from English to Mentalese. Since we understand Mentalese perfectly, the translation would be from an external spoken language into a language we understand. We could then rely on Mentalese, the true and only Language of Thought, to imbue English with meaning.

Readers may think that I merely make things more complicated by adding some mysterious new language to the mix. How can an unseen Language of Thought be empirically tested; isn't there a risk of this being a nontheory with no real empirical import? And what is the word *true* in (8) doing there?

Let's start with the word *true*. Saying what *true* actually means is so difficult as to be well beyond my abilities, but we can rely on the basic intuition that part of knowing the meaning of a sentence is knowing what the world would be like if the sentence were true. We could give a mathematical theory of truth, a theory that lays out how a sentence (or expression in Mentalese) could be true about the world. This might not work as a metaphysical definition of "the true" (whatever that is), but it could say something about the relation between language and the world and, in consequence, how language can carry information about the world.

Suppose I tell you something like

(9) I have a cocker spaniel named Sami.

You have several pieces of information from my utterance. Among other things, you know that I have something called a cocker spaniel and that this particular cocker spaniel answers to the name *Sami*. Of course, you can bring other information to bear on my statement; for example, you might also know that cocker spaniels are a kind of dog.

Now, what I said that I have a cocker spaniel named Sami is combined with what you already know that cocker spaniels are a kind of dog to *entail* that Sami is a dog and that I have a dog. This notion of *entailment* is defined in terms of truth and is important for understanding things like inference, the ability to combine bits of information to get new bits of information. Entailment can be defined as follows; I've simplified the definition somewhat, but it should be serviceable for now.

(10) ***Entailment***
A set of sentences $\{S_0, \ldots, S_i\}$ entails another sentence S_m if and only if sentence S_m must be true whenever all the sentences in $\{S_0, \ldots, S_i\}$ are simultaneously true.

Famous examples of entailment are Aristotelian syllogisms like

All men are mortal.
Socrates is a man.

Therefore, Socrates is mortal.

While the syllogisms may seem remote from experience, in fact entailment is used all the time in reasoning about the world. So the notion of truth in (8) is actually an important element in understanding how we use meaning in everyday language. Here we can begin to see a foundational fact for any theory of meaning: we use meanings to do things. Meanings are not simply an assemblage of facts; instead, they are tools for organizing behavior and thought; they are tools for operating on the physical and social world.

Concepts, Mentalese, and the Informational Universe

The next question that comes up regarding (8) is why Σ is couched in terms of Mentalese. Mentalese is a language of concepts. We all live in a physical world buzzing with clouds of particles, radiation of various sorts, and the interplay of fundamental forces. But that isn't the world of our experience. When I look around right now, I see my computer, my desk, a bunch of books, my telephone, my coffee cup, and so on. But these objects are informational things, not fundamental categories of the physical universe. Although a chair is a physical object, its role as a chair involves information; it requires that we recognize its function as a chair.

Where do these informational categories come from? Where does the informational universe come from? And given that we have these informational things, how would they be used in the real world?

My coffee cup must have certain physical properties to work as a coffee cup, but whether it's a coffee cup or not is largely up to me. I could use it as a paperweight, or as a shaving mug, or as a hat, or as a Christmas tree ornament, or as a collar ornament for my dog Sami. The role of my coffee cup in the world is only partly a matter of its physical properties. It has to work as a coffee cup for containing hot liquid, but its role is largely determined by how I fit its use into a broader scheme of things; it's really only a coffee cup if I decide to use it as such. I am the captain of my coffee cup. My mind makes it what it is. (I don't really believe this as stated, but let's go with it for the moment and work out what I could possibly be thinking.) In fact, I might decide that just about any receptacle for liquids could act as a coffee cup. After all, don't billionaires drink champagne

from ladies' slippers? What's to stop me from dubbing my shoe a coffee cup and drinking from it?

"Well," you might say, "that may hold for coffee cups and other things that people make. They're artifacts, and their use is a matter of human agency. But what about objects in the natural world, things that aren't artifacts." So let's take the case of a biological category, like "tiger." Is there some obvious physical property of tigers that make them, and nothing else, tigers?[2]

Tigers are striped quadrupeds that engage in predatory behavior; they have whiskers and big sharp teeth and claws. I might add that they are felines, but that category only makes sense inside a theory of biology, so let's set it aside for the moment.

None of the physical properties I mentioned are actually necessary criteria for tigerness. Suppose I had a tiger, Claude. He's a big striped quadruped, and he spends his time hunting. He indeed has big teeth and claws, very sharp and dangerous. So he's a tiger, as described.

But now suppose I take Claude to a laser hair removal center and have all his fur and whiskers lasered off. Claude's bald now, so he's no longer striped but he's still a tiger, near as anyone can tell.

What if I take Claude to a physical therapist who teaches him to walk (all the time) on his hind legs. He's still a tiger, even if he's no longer a quadruped.

Now suppose Claude has a spiritual conversion: predation and meat eating are wrong. From now on, Claude renounces meat and decides to eat grass. To accomplish this, he has his sharp teeth removed and special flat dentures installed, so that he can better grind up the grass with his new teeth. Furthermore, to ensure his pacific ways, he has his claws removed (perhaps I should call him de-Claude). Is he still a tiger? Yes, although at this point he looks and acts nothing like a tiger.

Perhaps Claude is a tiger because he has tiger DNA. But certainly the concept of "tiger" doesn't rely on DNA; after all, the concept of "tiger" was around before anyone had heard of DNA. People are mostly essentialists, I think. Once Claude has been fit into the concept of "tiger," he's treated essentially as a tiger, no matter what his appearance is. It's hard to get him out of that concept unless we open poor Claude up and discover that he was, all along, a robot. Then he becomes a "robot tiger," I suppose.[3]

Although it no doubt has some support from the physical world, perhaps "tiger" is largely an informational category. That category is as much about how we think about the world as it is about the physical

properties of the world; it seems that we have found the Language of Thought made manifest in the world.

Another example of the interaction of mind and world is our sense of number. Right now, I have four books on my desk, next to my computer. There is some evidence that I perceive the number four as an independent category; that is, part of my brain is devoted to the direct perception of number. Numbers may exist independently of our minds, or they may be constructed by us, but there can be no doubt that there is a specific neurobiological structure devoted to the perception of number. It's hard to see any physical world constant that would correlate with fourness. Nevertheless, we are able to extract numbers from the environment when called upon to do so. It seems as though number sense is a conceptual system that exists by virtue of the structure of our brains.

Our day-to-day talk is larded with all sorts of informational categories that have a very tenuous relation to the physical world. Suppose, watching the stock market fluctuate wildly in light of the crash in credit markets, I utter the following:

(11) The proposition that the invisible hand of the free market converts individual greed into social good is fundamentally flawed.

Surely, no one would expect what I said in (11) to be transparently supported by the physical world. It is riddled with concepts like "the invisible hand," "greed," and "social good," none of which have any transparent relation to physics.

This is really a very old point. In the first half of the twentieth century, a group of philosophers, the logical positivists, thought they could replace loose talk expressed in terms of abstract categories with precise talk grounded in physical measurement. The movement was short-lived, although quite influential; it didn't take long to realize that abstract categories are indispensable to our understanding of the world.

Language and the World

Of course, we're not free to treat concepts in any way we please. Concepts have to be tied to the world somehow. Example (8) showed a translation of a sentence *S* of an external language into a sentence Σ of Mentalese using a truth statement:

(12) S is true $\Leftrightarrow \Sigma$.

But this translation must be supported by another translation,

(13) Σ is true ⇔ TC,

where TC is a specification of the conditions under which the Mentalese sentence Σ holds true. That is, Mentalese, if it is to be useful at all, must be what is called an interpreted language that connects to the world. It needs to be interpreted because we cannot take its terms and predicates as basic. If we did take Mentalese as basic, a primitive, then our spoken language the language being interpreted by Mentalese would be unable to convey information about the world. But the fact that I made it to Prague, not Cleveland, shows that my language does have a connection with the world; if I'm to operate in the world and use language to learn about it and negotiate it, there must be a connection to the world.

In short, although language may translate to concepts, these concepts must relate to the world. We know this because we're able to coordinate our actions in the world using language. This means that Mentalese should be interpreted relative to a world model that is considered external to the speakers of a language:

Language → Mentalese → World model.

The world model would concern more than just the physical world; it would include abstract things like number and time, for example. But, crucially, it wouldn't be an internal, private representation of the world. It would be a shared public space, available to all speakers, that could be used to coordinate their verbal and conceptual behaviors. That way, when I say "dog" or "coffee cup" or "Prague," the corresponding concepts in Mentalese DOG or COFFEE CUP or PRAGUE would pick out dogs and coffee cups and Prague in the world model.

Platonic Heaven in a Box

Now, you might object that I just added more work. English must now be interpreted relative to Mentalese, and Mentalese relative to some model of the world. The following would doubtless be easier:

Language → World model.

Just skip the middleman and go directly to the world. I have some sympathy for this position, but let me try to give an answer that's fair to the Mentalese theorist.

We need a theory of linguistic meaning that properly connects language to both human reasoning and human action. Mentalese would be a common cognitive language that could connect these disparate areas

and organize them relative to the world. Furthermore, our concepts could be part of the world model, providing a way of making our private thoughts and opinions public. All this would be much harder to do without the common, mind-internal language of Mentalese.

Mentalese, of course, can't be exactly like a natural language. It's a language of mental representation that everyone uses but no one speaks. To make Mentalese work, we all need to have the same Mentalese concepts and agree as to how these Mentalese concepts pick out things in the world model, the simulacrum of the real world. This is a pretty tall order.

We can get some handle on the problem by consulting Plato's dialogue *Cratylus*. Hermogenes accosts Socrates and asks his help in solving a problem. Cratylus, the teacher of Hermogenes, teaches that there is a right and wrong way to call things. That is, each thing has a unique correct description, according to Cratylus. Protagoras, another teacher, claims that "man is the measure of all things." That is, there is no unique right or wrong way to call things: I use *dog* and the French use *chien*, and that's just the way it is. Neither of us is uniquely right; we're both right. Hermogenes wants Socrates to declare who is right: Cratylus or Protagoras.

The argument between Cratylus and Protagoras is really about whether linguistic signs are conventional (Protagoras's position) or natural (Cratylus's position, with which Socrates agrees). Note that whatever the signs of Mentalese are, they can't be conventional. Conventional things are arrived at through public practice, and there is nothing public about the signs of Mentalese; they're entirely internal to the brain or mind. We can only see Mentalese signs indirectly by virtue of our language use.

Early in the dialogue, Socrates lays the groundwork for his case that signs are natural:

Socrates But how about truth, then? You would acknowledge that there is in words a true and a false?
Hermogenes Certainly.
Socrates And there are true and false propositions?
Hermogenes To be sure.
Socrates And a true proposition says that which is, and a false proposition says that which is not?
Hermogenes Yes, what other answer is possible?
Socrates Then in a proposition there is a true and false?
Hermogenes Certainly.

Socrates But is a proposition true as a whole only, and are the parts untrue?
Hermogenes No, the parts are true as well as the whole.
Socrates Would you say the large parts and not the smaller ones, or every part?
Hermogenes I should say that every part is true.
Socrates Is a proposition resolvable into any part smaller than a name?
Hermogenes No, that is the smallest.
Socrates Then the name is a part of the true proposition?
Hermogenes Yes.
Socrates Yes, and a true part, as you say.
Hermogenes Yes.
Socrates And is not the part of a falsehood also a falsehood?
Hermogenes Yes.
Socrates Then, if propositions may be true and false, names may be true and false?
Hermogenes So we must infer.

In other words, a true sentence will be true in virtue of the truth of each and every one of its constituent parts. This passage anticipates an important idea in linguistics and the philosophy of language:

(14) ***Compositionality***
The meaning of a phrase is a function of the meanings of its parts and their mode of combination.

This is an extremely plausible idea that accounts for how each of us is capable of understanding new sentences. According to compositionality, I need to know the meanings of the atomic parts of the sentence, say, individual words, and I need to know how they combine to make up the whole sentence. That is, if I know what the words mean and I have a grammar that tells me how to combine words into sentences, then I can work out the meanings of sentences.

It is clear where Socrates is going with this argument. If a sentence is true, it must be because its parts are true. If the parts are true, it must be because their parts are true. And so on, down to the atomic level of words. It must be, then, that words are true of the objects they denote.

According to Socrates, there is a right and proper name for each thing, such name given by an "artificer of names" or "legislator" who skillfully associates with each thing the name it should have by nature:

Socrates Then, Hermogenes, I should say that the giving of names can be no such light matter as you fancy, or the work of light or chance

persons. And Cratylus is right in saying that things have names by nature, and that not every man is an artificer of names, but he only who looks to the name which each thing by nature has, and is able to express the true forms of things in letters and syllables.

There follows a lot of fanciful Greek etymology, designed to get at the true nature of things.

I doubt that many people would defend the natural theory of names that Socrates and Plato advance. It goes well with the idea of a Platonic heaven, where true forms dwell. Certainly, few would want to say that Greek or French or English words are more natural than those in another language. Everyone agrees that words are arbitrary symbols.

But what about the symbols of Mentalese? Mentalese is not supposed to vary in the way that natural languages vary. Everyone must be equipped with the same Mentalese.

A Mentalese theoretician would, I think, have to agree with Socrates that the signs of Mentalese are natural, not conventional. He would argue, I think, that the "artificer of names" is none other than evolution. Evolutionary psychology, which seeks to explain aspects of mind in terms of evolutionary theory, holds that we've evolved to have certain organs of perception, to act in certain ways in the world, and to think of the world in particular ways. Presumably, the way we think, perceive, and act has been of benefit to our species, aiding survival and reproduction. Hominid A, equipped with proto-Mentalese, is able to categorize and conceptualize the world in a useful way. She is better able to reason from the information she perceives. This adds to her reproductive success so that she passes on proto-Mentalese to her offspring. Hominid B is a clod with no internal representational capacity. He can't efficiently categorize or reason about the world. Being an ignorant oaf, he lacks hominid A's survival edge and is doomed.

Eventually, hominid A's proto-Mentalese would be passed on and modified into Mentalese. Mentalese itself, if it exists, would have to be part of our biological endowment. In other words, each of us would have to be born with an innate representational system that underlies our reasoning and action, the Language of Thought.

Inferences and Mentalese

I have some doubts about whether Mentalese predicates are heritable traits, but let's take a concrete example. Everybody has the concept

of causation as part of their internal representational system. Suppose there's a Mentalese expression, CAUSE, that we're all born with. It would work as follows: an AGENT would CAUSE an EVENT to transpire. In Mentalese,

(15) (CAUSE(EVENT))(AGENT).

Equally, we all have the notion, as part of our innate endowment, that things die, so Mentalese would include DIE. The thing that dies is not an AGENT; call it a THEME. The Mentalese expression would be

(16) DIE(THEME).

Now, we would also know

(17) DIE is a kind of EVENT.

We would learn that the English word *kill* means that the AGENT of *kill* caused the PATIENT to die. Thus,

(18) AGENT kill PATIENT ⇔ (CAUSE(DIE(PATIENT)))(AGENT).

Putting all this together, when a speaker of English hears

(19) John killed Bill.

she would translate it to the Mentalese expression

(20) (CAUSE(DIE(BILL))(JOHN),

where JOHN is the Mentalese symbol for John, and BILL is the Mentalese symbol for Bill. Because Mentalese is interpreted relative to a world model, she would know that John caused Bill to die in the world.

Even better, as an innately endowed speaker of Mentalese and a competent speaker of English, she might have access to the following rule:

(21) If (CAUSE(EVENT))(AGENT) then EVENT is true.

The rule in (21) is called a *meaning postulate*. It places a constraint on how causation is interpreted; if an event is caused, then that event actually has to happen.

Thus, we have the following translation from English to Mentalese:

(22) "John killed Bill" is true ⇔ (CAUSE(DIE(BILL)))(JOHN).

We also have the following correspondence from Mentalese to the world model:

(23) "(CAUSE(DIE(BILL)))(JOHN)" is true ⇔ John actually caused Bill to die in the world model.

Armed with the meaning postulate in (21), we can conclude that if John killed Bill, then Bill is dead. But this is an example of entailment. So this system of translations and meaning postulates actually can support an account of how we might reason with language.

When I was a boy, I had settled on the idea of English as Mentalese. It seemed utterly natural to me that *dog* meant dog and regrettable that Spanish speakers had to translate dog to *perro.*

Still, sometimes I would lie out on the grass in the backyard, watch the clouds, and repeat to myself "dog . . . dog . . . dog . . ." until the word itself disintegrated into just so much sonic nonsense. Then, the connection between *dog* and dog became mysterious, something to be wondered at. Why, I wondered, would *dog* mean dog?

And therein lies a problem. Suppose I could have repeated the Mentalese predicate DOG to myself. Is its connection to an actual dog any sturdier than the connection between *dog* and dog? The great artificer of names seems powerless here; how did I connect my mind-internal concept of DOG with that dog out in the real world?

Further Reading

A good place to start reading about truth is Blackburn's *Truth: A Guide* (2005). The translation theory of meaning is discussed in Quine's *Word and Object* (1960), and is critiqued in an article by Davidson (1974). The liar sentence in (5) is well-known; the multiple-sentence liar in (6) is adapted from Gupta and Belnap (1993). The true master of the liar paradox is Raymond Smullyan. His puzzle books are an encyclopedia of self-reference, but his masterwork is Smullyan (2009), which provides a kind of logical cosmology of lying and truth telling

Jerry Fodor has been an articulate champion of the Language of Thought; see Fodor (1975). I remember going to hear him as an undergraduate and being impressed when in response to a question from the audience, he argued that Neanderthals had the concept of "carburetor" as part of their innate Language of Thought. It's worth reading Fodor in tandem with Cowie's (1999) book, which gives a balanced discussion of nativism.

The translation statement in (8) is a deliberate conflation of an idea from Tarski (1983), who gave a mathematical definition of truth in formalized languages like logic. The idea is to transfer Tarski's approach to the Language of Thought.

A good discussion of number sense can be found in Dehaene (1997). Murray Grossman, a neurologist at the University of Pennsylvania, and I have worried about the relation between language and number sense; see Clark and Grossman (2007) for an interim report on the neurobiological underpinnings of language and number.

A good discussion of logical positivism and its downfall can be found in Soames (2003). Ray Jackendoff and Steven Pinker are both ardent defenders of Mentalese within linguistics. Fodor famously wrote a paper called "Three Reasons for Not Deriving 'Kill' from 'Cause to Die'" (1970), so he would surely not endorse my Mentalese analysis of *kill*. I certainly don't want to tar him with the brush of lexical decomposition (his theory is much more subtle). Nevertheless, the particular decompositional theory of meaning I described has wide currency in linguistics. For a very sophisticated version, see Hale and Keyser (2002) and the references cited there. See Jackendoff (1983) for a clear statement of Jackendoff's views. Pinker (1994) provides a widely read, very accessible discussion of generative grammar along with Mentalese. His more recent (2007) book delves into Mentalese and the structure of the lexicon.

Compositionality is often attributed to the nineteenth-century logician Gottlob Frege, although he didn't spell out exactly what he meant. See Dummett (1981) for some discussion.

A thorough discussion of inferencing and entailment can be found in any good introduction to logic. I'm particularly fond of the introductory text by Barwise and Etchemendy (1989).