distinction between what's innate and what's not. Clearly, everybody is going to put this line somewhere. For example, nobody is likely to think that the concept BROWN COW is primitive since, on the face of it, BROWN COW has BROWN and COW as constituents. Correspondingly, nobody is likely to think that the concept BROWN COW is innate since, on the face of it, it could be learned by being assembled from the previously mastered concepts BROWN and COW.

A lot of people have Very Strong Feelings about what concepts are allowed to be innate,[3] hence about how big a primitive conceptual basis an acceptable version of RTM can recognize. Almost everybody is prepared to allow RED in, and many of the liberal-minded will also let in CAUSE or AGENT. (See, for example, Miller and Johnson-Laird 1978). But there is, at present, a strong consensus against, as it might be, DOORKNOB or CARBURETTOR. I have no desire to join in this game of pick and choose since, as far as I can tell, it hasn't any rules. Suffice it that it would be nice if a theory of concepts were to provide a principled account of what's in the primitive conceptual basis, and it would be nice if the principles it appealed to were to draw the distinction at some independently plausible place. (Whatever, if anything, that means.) Chapter 6 will constitute an extended reconsideration of this whole issue, including the question just how the relation between a concept's being primitive and its being innate plays out. I hope there to placate such scruples about DOORKNOB and CARBURETTOR as some of you may feel, and to do so within the framework of an atomistic RTM.

5. Concepts are *public*; they're the sorts of things that lots of people can, and do, *share*.

Since, according to RTM, concepts are symbols, they are presumed to satisfy a type/token relation; to say that two people share a concept (i.e. that they have literally the same concept) is thus to say that they have tokens of literally the same concept type. The present requirement is that the conditions for typing concept tokens must not be so stringent as to assign practically every concept token to a different type from practically any other.

[3] I put it this way advisedly. I was once told, in the course of a public discussion with an otherwise perfectly rational and civilized cognitive scientist, that he "could not permit" the concept HORSE to be innate in humans (though I guess it's OK for it to be innate in horses). I forgot to ask him whether he was likewise unprepared to permit neutrinos to lack mass.

Just why feelings run so strongly on these matters is unclear to me. Whereas the ethology of all other species is widely agreed to be thoroughly empirical and largely morally neutral, a priorizing and moralizing about the ethology of our species appears to be the order of the day. Very odd.

It seems pretty clear that all sorts of concepts (for example, DOG, FATHER, TRIANGLE, HOUSE, TREE, AND, RED, and, surely, lots of others) are ones that all sorts of people, under all sorts of circumstances, have had and continue to have. A theory of concepts should set the conditions for concept possession in such a way as not to violate this intuition. Barring very pressing considerations to the contrary, it should turn out that people who live in very different cultures and/or at very different times (me and Aristotle, for example) both have the concept FOOD; and that people who are possessed of very different amounts of mathematical sophistication (me and Einstein, for example) both have the concept TRIANGLE; and that people who have had very different kinds of learning experiences (me and Helen Keller, for example) both have the concept TREE; and that people with very different amounts of knowledge (me and a four-year-old, for example) both have the concept HOUSE. And so forth. Accordingly, if a theory or an experimental procedure distinguishes between my concept DOG and Aristotle's, or between my concept TRIANGLE and Einstein's, or between my concept TREE and Helen Keller's, etc. that is a very strong prima facie reason to doubt that the theory has got it right about concept individuation or that the experimental procedure is really a measure of concept possession.

I am thus setting my face against a variety of kinds of conceptual relativism, and it may be supposed that my doing so is itself merely dogmatic. But I think there are good grounds for taking a firm line on this issue. Certainly RTM is required to. I remarked in Chapter 1 that RTM takes for granted the centrality of intentional explanation in any viable cognitive psychology. In the cases of interest, what makes such explanations intentional is that they appeal to covering generalizations about people who believe that such-and-such, or people who desire that so-and-so, or people who intend that this and that, and so on. In consequence, the extent to which an RTM can achieve generality in the explanations it proposes depends on the extent to which mental contents are supposed to be shared. If everybody else's concept WATER is different from mine, then it is literally true that only I have ever wanted a drink of water, and that the intentional generalization 'Thirsty people seek water' applies only to me. (And, of course, only I can state that generalization; words express concepts, so if your WATER concept is different from mine, 'Thirsty people seek water' means something different when you say it and when I do.) Prima facie, it would appear that any very thoroughgoing conceptual relativism would preclude intentional generalizations with any very serious explanatory power. This holds in spades if, as seems likely, a coherent conceptual relativist has to claim that conceptual identity can't be maintained even across time slices of the same individual.

There is, however, a widespread consensus (and not only among conceptual relativists) that intentional explanation can, after all, be preserved without supposing that belief contents are often—or even ever—literally public. The idea is that a robust notion of content *similarity* would do just as well as a robust notion of content *identity* for the cognitive scientist's purposes. Here, to choose a specimen practically at random, is a recent passage in which Gil Harman enunciates this faith:

> Sameness of meaning from one symbol system to another is a similarity relation rather than an identity relation in the respect that sameness of meaning is not transitive . . . I am inclined to extend the point to concepts, thoughts, and beliefs . . . The account of sameness of content appeals to the best way of translating between two systems, where goodness in translation has to do with preserving certain aspects of usage, with no appeal to any more 'robust' notion of content or meaning identity . . . [There's no reason why] the resulting notion of sameness of content should fail to satisfy the purposes of intentional explanation. (1993: 169–79)[4]

It's important whether such a view can be sustained since, as we'll see, meeting the requirement that intentional contents be literally public is non-trivial; like compositionality, publicity imposes a substantial constraint upon one's theory of concepts and hence, derivatively, upon one's theory of language. In fact, however, the idea that content similarity is the basic notion in intentional explanation is affirmed a lot more widely than it's explained; and it's quite unclear, on reflection, how the notion of similarity that such a semantics would require might be unquestion-beggingly developed. On one hand, such a notion must be robust in the sense that it preserves intentional explanations pretty generally; on the other hand, it must do so *without itself presupposing a robust notion of content identity*. To the best of my knowledge, it's true *without exception* that all the construals of concept similarity that have thus far been put on offer egregiously fail the second condition.

Harman, for example, doesn't say much more about content-similarity-cum-goodness-of-translation than that it isn't transitive and that it "preserves certain aspects of usage". That's not a lot to go on. Certainly it leaves wide open whether Harman is right in denying that his account of content similarity presupposes a "'robust' notion of content or meaning identity". For whether it does depends on how the relevant "aspects of

[4] See also Smith, Medin, and Rips: "what accounts for categorization cannot account for stability [publicity] . . . [a]s long as *stability of concepts* is equated with *sameness of concepts* . . . But there is another sense of stability, which can be equated with *similarity of mental contents* . . . and for this sense, what accounts for categorization may at least partially account for 'stability' "(1984: 268). Similar passages are simply ubiquitous in the cognitive science literature; I'm grateful to Ron Mallon for having called this example to my attention.

usage" are themselves supposed to be individuated, and about this we're told nothing at all.

Harman is, of course, too smart to be a behaviourist; 'usage', as he uses it, is itself an intentional-cum-semantic term. Suppose, what surely seems plausible, that one of the 'aspects of usage' that a good translation of 'dog' has to preserve is that it be a term that implies *animal*, or a term that doesn't apply to ice cubes, or, for matter, a term that means *dog*. If so, then we're back where we started; Harman needs notions like *same* implication, *same* application, and *same* meaning in order to explicate his notion of content similarity. All that's changed is which shell the pea is under.

At one point, Harman asks rhetorically, "What aspects of use determine meaning?" Reply: "It is certainly relevant what terms are applied to and the reasons that might be offered for this application . . . it is also relevant how some terms are used in relation to other terms" (ibid.: 166). But I can't make any sense of this unless some notion of 'same application', 'same reason', and 'same relation of terms' is being taken for granted in characterizing what good translations *ipso facto* have in common. NB on pain of circularity: *same* application (etc.), not *similar* application (etc.). Remember that *similarity of semantic properties* is the notion that Harman is trying to explain, so his explanation mustn't *presuppose* that notion.

I don't particularly mean to pick on Harman; if his story begs the question it was supposed to answer, that is quite typical of the literature on concept similarity. Though it's often hidden in a cloud of technical apparatus (for a detailed case study, see Fodor and Lepore 1992: ch. 7), the basic problem is easy enough to see. Suppose that we want the following to be a prototypical case where you and I have different but similar concepts of George Washington: though we agree about his having been the first American President, and the Father of His Country, and his having cut down a cherry tree, and so on, you think that he wore false teeth and I think that he didn't. The similarity of our GW concepts is thus some (presumably weighted) function of the number of propositions about him that we both believe, and the dissimilarity of our GW concepts is correspondingly a function of the number of such propositions that we disagree about. So far, so good.

But the question now arises: what about the shared beliefs themselves; are they or aren't they *literally* shared? This poses a dilemma for the similarity theorist that is, as far as I can see, unavoidable. If he says that our agreed upon beliefs about GW are literally shared, then he hasn't managed to do what he promised; viz. introduce a notion of similarity of content that dispenses with a robust notion of publicity. But if he says that the agreed beliefs aren't literally shared (viz. that they are only required to be similar), then his account of content similarity begs the very question it was supposed to answer: his way of saying what it is for concepts to have similar but not identical contents presupposes a prior notion of beliefs with similar but not identical contents.

The trouble, in a nutshell, is that all the obvious construals of *similarity of beliefs* (in fact, all the construals that I've heard of) take it to involve *partial overlap* of beliefs.[5] But this treatment breaks down if the beliefs that are in the overlap are themselves construed as similar but not identical. It looks as though a robust notion of content similarity *can't but* presuppose a correspondingly robust notion of content identity. Notice that this situation is not symmetrical; the notion of content identity doesn't require a prior notion of content similarity. Leibniz's Law tells us what it is for the contents of concepts to be identical; Leibniz's Law tells us what it is for *anythings* to be identical.

As I remarked above, different theorists find different rugs to sweep this problem under; but, as far as I can tell, none of them manages to avoid it. I propose to harp on this a bit because confusion about it is rife, not just in philosophy but in the cognitive science community at large. Not getting it straight is one of the main things that obscures how very hard it is to construct a theory of concepts that works, and how very much cognitive science has thus far failed to do so.

Suppose, for example, it's assumed that your concept PRESIDENT is similar to my concept PRESIDENT in so far as we assign similar subjective probabilities to propositions that contain the concept. There are plenty of reasons for rejecting this sort of model; we'll discuss its main problems in Chapter 5. Our present concern is only whether constructing a probabilistic account of concept similarity would be a way to avoid having to postulate a robust notion of content identity.

Perhaps, in a typical case, you and I agree that p is very high for 'FDR is/was President' and for 'The President is the Commander-in-Chief of the Armed Forces' and for 'Presidents have to be of voting age', etc.; but, whereas you rate 'Millard Fillmore is/was President' as having a probability close to 1, I, being less well informed, take it to be around p = 0.07 (*Millard Fillmore*???). This gives us an (arguably) workable construal of the idea that we have similar but not identical PRESIDENT concepts. But it does so only by helping itself to a prior notion of belief identity, and to the assumption that there are lots of thoughts of which

[5] 'Why not take content similarity as primitive and *stop trying* to construe it?' Sure; but then why not take content *identity* as primitive and stop trying to construe *it*? In which case, what is semantics *for*?

our respective PRESIDENTs are constituents that we literally share. Thus, you and I are, by assumption, both belief-related to the thoughts that Millard Fillmore was President, that Presidents are Commanders-in-Chief, etc. The difference between us is in the *strengths* of our beliefs, not in their contents.[6] And, as usual, it really does seem to be *identity* of belief content that's needed here. If our respective beliefs about Presidents having to be of voting age were supposed to be merely *similar*, circularity would ensue: since content similarity is the notion we are trying to explicate, it mustn't be among the notions that the explication presupposes. (I think I may have mentioned that before.)

The same sort of point holds, though even more obviously, for other standard ways of construing conceptual similarity. For example, if concepts are sets of features, similarity of concepts will presumably be measured by some function that is sensitive to the amount of overlap of the sets. But then, the atomic feature assignments must themselves be construed as literal. If the similarity between your concept CAT and mine depends (*inter alia*) on our agreement that '+ has a tail' is in both of our feature bundles, then the assignment of that feature to these bundles must express a literal consensus; it must literally be the property of *having a tail* that we both literally think that cats literally have. (As usual, nothing relevant changes if feature assignments are assumed to be probabilistic or weighted; or if the feature assigned are supposed to be "subsemantic", though these red herrings are familiar from the Connectionist literature.)

Or, suppose that concepts are thought of as positions in a "multi-dimensional vector space" (see e.g. Churchland 1995) so that the similarity between your concepts and mine is expressed by the similarity of their positions in our respective spaces. Suppose, in particular, that it is constitutive of the difference between our NIXON concepts that you think Nixon was even more of a crook than I do. Once again, a robust notion of content identity is presupposed since each of our spaces is required to have a dimension that expresses crookedness; a fortiori, both are required

[6] Alternatively, a similarity theory might suppose that what we share when our PRESIDENT concepts are similar are similar beliefs about the probabilities of certain propositions: you believe that p(presidents are CICs) = 0.98; I believe that p(presidents are CICs) = 0.95; Bill believes that p(Presidents are CICs) = 0.7; so, all else equal, your PRESIDENT concept is more like mine than Bill's is.

But this construal does nothing to discharge the basic dependence of the notion of content similarity on the notion of content identity since what it says makes our beliefs similar is that they make similar estimates of the probability *of the very same proposition*; e.g. of the proposition that presidents are CICs. If, by contrast, the propositions to which our various probability estimates relate us are themselves supposed to be merely similar, then it does *not* follow from these premisses that *ceteris paribus* your PRESIDENT concept is more like mine than like Bill's.

to have dimensions which express degrees of *the very same property*. That should seem entirely unsurprising. Vector space models identify the dimensions of a vector space *semantically* (viz. by stipulating what the location of a concept along that dimension is to *mean*), and it's just a truism that the positions along dimension *D* can represent degrees of *D*-ness only in a mind that possesses the concept of being *D*. You and I can argue about whether Nixon was merely crooked or very crooked only if the concept of *being crooked* is one that we have in common.

It may seem to you that I am going on about such truisms longer than necessity demands. It often seems that to me, too. There are, however, at least a zillion places in the cognitive science literature, and at least half a zillion in the philosophy literature, where the reader is assured that some or all of his semantical troubles will vanish quite away if only he will abandon the rigid and reactionary notion of content identity in favour of the liberal and laid-back notion of content similarity. But in none of these places is one ever told how to do so. That's because nobody has the slightest idea how. In fact, it's all just loose talk, and it causes me to grind my teeth.

Please note that none of this is intended to claim that notions like belief similarity, content similarity, concept similarity, etc. play less than a central role in the psychology of cognition. On the contrary, for all I know (certainly for all I am prepared non-negotiably to assume) it may be that every powerful intentional generalization is of the form "If *x* has a belief similar to *P*, then . . ." rather than the form "If *x* believes *P*, then . . .". If that is so, then so be it. My point is just that assuming that it is so doesn't exempt one's theory of concepts from the Publicity constraint. To repeat one last time: all the theories of content that offer a robust construal of conceptual similarity do so by presupposing a correspondingly robust notion of concept identity. As far as I can see, this is unavoidable. If I'm right that it is, then the Publicity constraint is *ipso facto* non-negotiable.

OK, so those are my five untendentious constraints on theories of concepts. In succeeding chapters, I'll consider three stories about what concepts are; viz. that they are definitions; that they are prototypes/stereotypes; and (briefly) something called the 'theory theory' which says, as far as I can make out, that concepts are abstractions from belief systems. I'll argue that each of these theories violates at least one of the non-negotiable constraints; and that it does so, so to speak, not a little bit around the edges but egregiously and down the middle. We will then have to consider what, if any, options remain for developing a theory of concepts suitable to the purposes of an RTM.

Before we settle down to this, however, there are a last couple of preliminary points that I want to put in place.

Here is the first: although I'm distinguishing three theories of concepts for purposes of exposition and attack, and though supporters of each of these theories have traditionally wanted to distance themselves as much as possible from supporters of the others, still all three theories are really versions of one and the same idea about content. I want to stress this since I'm going to argue that it is primarily because of what they agree about that all three fail.

The theories of concepts we'll be considering all assume a metaphysical thesis which, as I remarked in Chapter 1, I propose to reject: namely, that primitive concepts, and (hence) their possession conditions, are at least partly constituted by their inferential relations. (That complex concepts—BROWN COW, etc.—and their possession conditions are exhaustively constituted by their inferential relations to their constituent concepts is not in dispute; to the contrary, compositionality requires it, and compositionality isn't negotiable.) The current near-universal acceptance of Inferential Role Semantics in cognitive science marks a radical break with the preceding tradition in theories about mind and language: pre-modern theories typically supposed that primitive concepts are individuated by their (e.g. iconic or causal) relations to things in the world. The history of the conversion of cognitive scientists to IR semantics would make a book by itself; a comedy, I think, though thus far without a happy ending:

—In philosophy, the idea was pretty explicitly to extend the Logicist treatment of logical terms into the non-logical vocabulary; if IF and SOME can be identified with their inferential roles, why not TABLE and TREE as well?

—In linguistics, the idea was to extend to semantics the Structuralist notion that a level of grammatical description is a 'system of differences': if their relations of equivalence and contrast are what bestow phonological values on speech sounds, why shouldn't their relations of implication and exclusion be what bestow semantic values on forms of words?

—In AI, the principle avatar of IRS was 'procedural semantics', a deeply misguided attempt to extend the principle of 'methodological solipsism' from the theory of mental processes to the theory of meaning: if a mental process (thinking, perceiving, remembering, and the like) can be 'purely computational' why can't conceptual content be purely computational too? If computers qua devices that perform inferences can *think*, why can't computers qua devices that perform inferences *mean*?

—I don't know how psychology caught IRS; perhaps it was from philosophy, linguistics, and AI. (I know one eminent developmental psychologist who certainly caught it from Thomas Kuhn.) Let that be an object lesson in the danger of mixing disciplines. Anyhow, IRS got to be the fashion in psychology too. Perhaps the main effect of the "cognitive revolution" was that espousing some or other version of IRS became the received way for a psychologist not to be a behaviourist.

So, starting around 1950, practically everybody was saying that the '"Fido"–Fido fallacy' is fallacious,[7] and that concepts (/words) are like chess pieces: just as there can't be a rook without a queen, so there can't be a DOG without an ANIMAL. Just as the value of the rook is partly determined by its relation to the queen, so the content of DOG is partly determined by its relation to ANIMAL. Content is therefore a thing that can only happen internal to *systems* of symbols (or internal to languages, or, on some versions, internal to forms of life). It was left to 'literary theory' to produce the *reductio ad absurdum* (literary theory is good at that): content is constituted *entirely* by intra-symbolic relations; just as there's nothing 'outside' the chess game that matters to the values of the pieces, so too there's nothing outside the text that matters to what it means. Idealism followed, of course.

It is possible to feel that these various ways of motivating IRS, historically effective though they clearly were, are much less than overwhelmingly persuasive. For example, on reflection, it doesn't seem that languages are a lot like games after all: queens and pawns don't mean anything, whereas 'dog' means *dog*. That's why, though you can't translate the queen into French (or, a fortiori, into checkers), you can translate 'dog' into 'chien'. It's perhaps unwise to insist on an analogy that misses so glaring a difference.

Phonemes don't mean anything either, so prima facie, *pace* Saussure, "having a phonological value" and "having a semantic value" would seem to be quite different sorts of properties. Even if it were right that phonemes are individuated by their contrasts and equivalences—which probably they aren't—that wouldn't be much of a reason to claim that words or concepts are also individuated that way.

If, in short, one asks to hear some serious arguments for IRS, one discovers, a bit disconcertingly, that they are very thin upon the ground. I think that IRS is most of what is wrong with current theorizing in cognitive science and the metaphysics of meaning. But I don't suppose for a minute that any short argument will, or should, persuade you to consider junking it. I expect that will need a long argument; hence this long book. Long arguments take longer than short arguments, but they do sometimes create conviction.

Accordingly, my main subject in what follows will be not the history of

[7] That is, the "fallacy" of assuming that the meaning of the word is the eponymous dog.

IR semantics, or the niceties of its formulation, or its evidential status, but rather its impact on empirical theories of concepts. The central consideration will be this: If you wish to hold that the content of a concept is constituted by the inferences that it enters into, you are in need of a principled way of deciding *which inferences constitute which concepts*. What primarily distinguishes the cognitive theories we'll consider is how they answer this question. My line will be that, though as far as anybody knows the answers they offer exhaust the options, pretty clearly none of them can be right. Not, NB, that they are incoherent, or otherwise confused; just that they fail to satisfy the empirical constraints on theories of concepts that I've been enumerating, and are thus, almost certainly, false.

At that point, I hope that abandoning IRS in favour of the sort of atomistic, informational semantics that I tentatively endorsed in Chapter 1 will begin to appear to be the rational thing to do. I'll say something in Chapter 6 about what this sort of alternative to IRS might be like.

So much for the first of my two concluding addenda. Here is the second:

I promised you in Chapter 1 that I wouldn't launch yet another defence of RTM; I proposed—aside from my admittedly tendentious endorsement of informational semantics—simply to take RTM for granted as the context in which problems about the nature of concepts generally arise these days. I do mean to stick to this policy. Mostly. But I can't resist rounding off these two introductory chapters by remarking how nicely the pieces fit when you put them all together. I'm going to exercise my hobbyhorse after all, but only a little.

In effect, in these introductory discussions, we've been considering constraints on a theory of cognition that emerge from two widely different, and largely independent, research enterprises. On the one hand, there's the attempt to save the architecture of a Fregean—viz. a purely referential—theory of meaning by taking seriously the idea that concepts can be distinguished by their 'modes of presentation' of their extensions. It's supposed to be modes of presentation that answer the question 'How can coreferential concepts be distinct?' Here Frege's motives concur with those of Informational Semantics; since both are referential theories of content, both need a story about how thinking about the Morning Star could be different from thinking about the Evening Star, given that the two thoughts are connected with the same 'thing in the world'.

The project of saving the Frege programme faces two major hurdles. First, 'Mates cases' appear to show that modes of presentations can't be senses. Frege to the contrary notwithstanding, it looks as though practically any linguistic difference between prima facie synonymous expressions, merely syntactic differences distinctly included, can be recruited to block their substitution in some Mates context or other. In the current jargon, the individuation of the propositional attitudes apparently slices them about as thin as the syntactic individuation of forms of words, hence not only thinner than reference can, *but also thinner than sense can.*

The other obstacle to saving the Frege programme was that it took for granted that the semantic question 'How can coreferential concepts fail to be synonyms?' gets the same answer as the psychological question 'How can there be more than one way of grasping a referent?' The postulation of senses was supposed to answer both questions. I argued, however, that given Frege's Platonism about senses, it's by no means obvious why his answer to the first would constitute an answer to the second; in effect, Frege simply stipulates their equivalence. I supposed the moral to be that Frege's theoretical architecture needs to be explicitly psychologized. Modes of presentation need to be 'in the head'.

The short form is: the Frege programme needs something that is both in the head and of the right kind to distinguish coreferential concepts, and the Mates cases suggest that whatever is able to distinguish coreferential concepts is apt for syntactic individuation. Put all this together and it does rather suggest that modes of presentation are syntactically structured mental particulars. Suggestion noted.

The other research programme from which my budget of constraints on theories of concepts derived is the attempt, in cognitive science, to explain how a finite being might have intentional states and capacities that are productive and systematic. This productivity/systematicity problem again has two parts: 'Explain how there can be infinitely many propositional attitudes each with its distinctive propositional object (i.e. each with its own content)' and: 'Explain how there can be infinitely many propositional attitudes each with its distinctive causal powers (i.e. each with its own causal role in mental processes).' Here I have followed what Pylyshyn and I (Fodor and Pylyshyn 1988) called the 'Classical' computational tradition that proceeds from Turing: mental representations are syntactically structured. Their conditions of semantic evaluation and their causal powers both depend on their syntactic structures; the former because mental representations have a compositional semantics that is sensitive to the syntactic relations among their constituents; the latter because mental processes are *computations* and are thus syntactically driven by definition. So the Classical account of productivity/systematicity points in much the same direction as the psychologized Frege programme's account of the individuation of mental states: viz. towards syntactically structured mental particulars whose tokenings are matched, case for case, with tokenings of the de dicto propositional attitudes.

Syntactically structured mental particulars whose tokenings are matched, case for case, with tokenings of the de dicto propositional attitudes are, of course, exactly what RTM has for sale. So RTM seems to be what both the Frege/Mates problems and the productivity/systematicity problems converge on. If beliefs (and the like) are relations to syntactically structured mental representations, there are indeed two parameters of belief individuation, just as Frege requires: Morning Star beliefs have the same conditions of semantic evaluation as Evening Star beliefs, but they implicate the tokening of different syntactic objects and are therefore different beliefs with different causal powers. That believing *P* and believing *Q* may be different mental states even if '*P*' and '*Q*' have the same semantic value shows up in the Mates contexts. That believing *P* and believing *Q* may have different causal powers even if '*P*' and '*Q*' have the same semantic value shows up in all those operas where the soprano dies of mistaken identity.

So RTM looks like a plausible answer to several questions that one might have supposed to be unrelated. I hope that isn't an accident. This book runs on the assumption that it isn't, hence that we need RTM a lot. RTM, in turn, needs a theory of concepts a lot since compositionality says that the contents and causal powers of mental representations are both inherited, eventually, from the contents and causal powers of their minimal constituents; viz. from the primitive concepts that they contain. RTM is simply *no good* without a viable theory of concepts.

So be it, then. Let's see what there might be on offer.

3

The Demise of Definitions, Part I: The Linguist's Tale

> Certain matters would appear to get carried certain distances whether one wishes them to or not, unfortunately.
>
> —David Markham, *Wittgenstein's Mistress*

Introduction

I WANT to consider the question whether concepts are definitions. And let's, just for the novelty, start with some propositions that are clearly true:

1. You can't utter the expression 'brown cow' without uttering the word 'brown'.
2. You can utter the word 'bachelor' without uttering the word 'unmarried'.

The asymmetry between 1 and 2 will be granted even by those who believe that the "semantic representation" of 'bachelor' (its representation, as linguists say, "at the semantic level") is a complex object which contains, *inter alia*, the semantic representation of 'unmarried'.

Now for something that's a little less obvious:

3. You can't entertain the M(ental) R(epresentation) BROWN COW without entertaining the MR BROWN.
4. You can't entertain the M(ental) R(epresentation) BACHELOR without entertaining the MR UNMARRIED.

I'm going to take it for granted that 3 is true. I have my reasons; they'll emerge in Chapter 5. Suffice it, for now, that anybody who thinks that 3 and the like are false will certainly think that 4 and the like are false; and that 4 and the like are indeed false is the main conclusion this chapter aims at. I pause, however, to remark that 3 is meant to be tendentious. It claims not just what everyone admits, viz. that anything that satisfies BROWN COW *inter alia* satisfies BROWN, viz. that brow cows are *ipso facto* brown.

Proposition 3 says, moreover, that to think the content *brown cow* is, *inter alia*, to think the concept BROWN, and that would be false if the mental representation that expresses *brown cow* is atomic; like, for example, BROWNCOW.

What about 4? Here again there is a way of reading what's being claimed that makes it merely truistic: viz. by not distinguishing *concept* identity from *content* identity. It's not, I suppose, unreasonable (for the present illustrative purposes, I don't care whether it's true) to claim that the content *bachelor* and the content *unmarried man* are one and the same. For example, if concepts express properties, then it's not unreasonable to suppose that BACHELOR and UNMARRIED MAN express the *same* property. If so, and if one doesn't distinguish between content identity and concept identity, then of course it follows that you can't think BACHELOR without thinking UNMARRIED (unless you can think UNMARRIED MAN without thinking UNMARRIED. Which let's just concede that you can't).[1]

However, since we *are* distinguishing content identity from concept identity, we're not going to read 4 that way. Remember that RTM is in force, and RTM says that to each tokening of a mental state with the content *so-and-so* there corresponds a tokening of a mental representation with the content *so-and-so*. In saying this, RTM explicitly means to leave open the possibility that different (that is, type distinct) mental representations might correspond to the same content; hence the analogy between mental representations and modes of presentation that I stressed in Chapter 2. In the present case, the concession that being a bachelor and being an unmarried man are the same thing is meant to leave open the question whether BACHELOR and UNMARRIED MAN are the same concept.

RTM also says that (infinitely many, but not all) mental representations have constituent structure; in particular that there are both complex

[1] It will help the reader to keep the uses distinct from the mentions, to bear in mind that the expressions appearing in caps. (e.g. 'BACHELOR') are *names*, rather than *structural descriptions*, of mental representations. I thus mean to leave it open that the MR that 'BACHELOR' names might be structurally complex; for example, it might have as constituents the MRs that 'UNMARRIED' and 'MAN' name. By contrast, it's stipulative that no formula is a *structural description* of a mental representation unless it contains names of the MR's constituents. The issues we'll be concerned with can often be phrased either by asking about the structure of mental representations or about the structural descriptions of mental representations. In practice, I'll go back and forth between the two.

The claim that concepts are definitions can be sharpened in light of these remarks. Strictly speaking, it's that the *definiens* is the structural description of the *definiendum*; for example, 'UNMARRIED MAN' is the structural description of the concept BACHELOR.

mental representations and primitive mental representations, and that the former have the latter as proper parts. We are now in a position to make expository hay out of this assumption; we can rephrase the claim that is currently before the house as:

5. The M(ental) R(epresentation) UNMARRIED, which is a constituent of the MR UNMARRIED MAN, is likewise a constituent of the MR BACHELOR.

Here's a standard view: the concept BACHELOR is expressed by the word "bachelor", and the word "bachelor" is definable; it means the same as the phrase "unmarried man". In the usual case, the mental representation that corresponds to a concept that corresponds to a definable word is complex: in particular, the mental representation that corresponds to a definable word usually has the same constituent structure as the mental representation that corresponds to its definition. So, according to the present proposal, the constituent structure of the mental representation BACHELOR is something like 'UNMARRIED MAN'.

The thesis that definition plays an important role in the theory of mental representation will be the main concern in this chapter and the next. According to that view, many mental representations work the way we've just supposed that BACHELOR does. That is, they correspond to concepts that are expressed by definable words, and they are themselves structurally complex. This thesis is, to put it mildly, *very* tendentious. In order for it to be true, it must turn out that there are many definable words; and it must turn out, in many cases, that the MRs that correspond to these definable words are structurally complex. I'm going to argue that it doesn't, in fact, turn out in either of those ways.[2]

One last preliminary, and then we'll be ready to go. If there are no definable words, then, of course, there are no complex mental representations that correspond to them. But it doesn't follow that if there are many complex mental representations, then lots of words are definable. In fact, I take it that the view now favoured in both philosophy and cognitive science is that most words aren't definable but do correspond to

[2] It's common ground that—idioms excepted—MRs that correspond to phrases (for example, the one that corresponds to "brown cow") are typically structurally complex, so I've framed the definition theory as a thesis about the MRs of concepts that are expressed by lexical items. But, of course, this way of putting it relativizes the issue to the choice of a reference language. Couldn't it be that the very same concept that is expressed by a single word in English gets expressed by a phrase in Bantu, or vice versa? Notice, however, that this could happen only if the English word in question is definable; viz. definable in Bantu. Since it's going to be part of my story that most words are undefinable—not just undefinable in the language that contains them, but undefinable *tout court*—I'm committed to claiming that this sort of case can't arise (very often). The issue is, of course, empirical. So be it.

complex MRs (to something like prototypes or exemplars). Since the case against definitions isn't *ipso facto* a case against complex mental representations, I propose the following expository strategy. In this chapter and the next, I argue that concepts *aren't* definitions even if lots of mental representations *are* complex. Chapter 5 will argue that there are (practically) no complex mental representations at all, definitional or otherwise.[3] At that point, atomism will be the option of last resort.

If we thus set aside, for the moment, all considerations that don't distinguish the claim that mental representations are typically definitional from the weaker claim that mental representations are typically complex, what arguments have we left to attend to? There are two kinds: the more or less empirical ones and the more or less philosophical ones. The empirical ones turn on data that are supposed to show that the mental representations that correspond to definable words are, very often and simply as a matter of fact, identical to the mental representations that correspond to phrases that define the words. The philosophical ones are supposed to show that we need mental representations to be definitions because nothing else will account for our intuitions of conceptual connectedness, analyticity, a prioricity, and the like. My plan is to devote the rest of this chapter to the empirical arguments and all of Chapter 4 to the philosophical arguments. You will be unsurprised to hear what my unbiased and judicious conclusion is going to be. My unbiased and judicious conclusion is going to be that neither the philosophical nor the empirical arguments for definitions are any damned good.

So, then, to business.

Almost everybody used to think that concepts are definitions; hence that having a concept is being prepared to draw (or otherwise acknowledge) the inferences that define it. Prima facie, there's much to be said for this view. In particular, definitions seem to have a decent chance of satisfying all five of the 'non-negotiable' conditions which Chapter 2 said that concepts have to meet. If the meaning-constitutive inferences are the defining ones, then it appears that:

—Definitions can be mental particulars if any concepts can. Whatever the definition of 'bachelor' is, it has the same ontological status as the mental representation that you entertain when you think *unmarried man*. That there is such a mental representation is a claim to which RTM is, of course, independently committed.

—Semantic evaluability is assured; since *all* inferences are semantically

[3] i.e. there are no complex mental representations other than those that correspond to concepts that are expressed by phrases; see the preceding footnote. From now on, I'll take this caveat for granted.

evaluable (for soundness, validity, reliability, etc.), defining inferences are semantically evaluable *inter alia*.

—Publicity is satisfied since there's no obvious reason why lots of people might not assign the same defining inferences to a given word or concept. They might do so, indeed, even if there are lots of differences in what they know/believe about the things the concept applies to (lots of differences in the 'collateral information' they have about such things).

—Compositionality is satisfied. This will bear emphasis later. I'm going to argue that, of the various 'inferential role' theories of concepts, only the one that says that concepts are definitions meets the compositionality condition. Suffice it for now that words/concepts do contribute their definitions to the sentences/thoughts that contain them; it's part and parcel of 'bachelor' meaning *unmarried man* that the sentence 'John is a bachelor' means *John is an unmarried man* and does so because it has 'bachelor' among its constituents. To that extent, at least, definitions are in the running to be both word meanings and conceptual contents.

—Learnability is satisfied. If the concept DOG is a definition, then learning the definition should be all that's required to learn the concept. A fortiori, concepts that are definitions don't have to be innate.

To be sure, learning definitions couldn't be the *whole* story about acquiring concepts. Not all concepts could be definitions, since some have to be the primitives that the others are defined in terms of; about the acquisition of the primitive concepts, some quite different story will have to be told. What determines which concepts are primitive was one of the questions that definition theories never really resolved. Empiricists in philosophy wanted the primitive concepts to be picked out by some epistemological criterion; but they had no luck in finding one. (For discussion of these and related matters, see Fodor 1981*a*, 1981*b*.) But, however exactly this goes, the effect of supposing that there are definitions is to reduce the problems about concepts at large to the corresponding problems about primitive concepts. So, if some (complex) concept *C* is defined by primitive concepts $c_1, c_2, \ldots$, then explaining how we acquire *C* reduces to explaining how we acquire $c_1, c_2, \ldots$ And the problem of how we apply *C* to things that fall under it reduces to the problem of how we apply $c_1, c_2, \ldots$ to the things that fall under them. And explaining how we reason with *C* reduces to explaining how we reason with $c_1, c_2, \ldots$ And so forth. So there is good work for definitions to do if there turn out to be any.

All the same, these days almost nobody thinks that concepts are definitions. There is now something like a consensus in cognitive science that the notion of a definition has no very significant role to play in theories of meaning. It is, to be sure, a weakish argument against