JOHN LOCKE
Portrait by John Greenhill, *c.* 1672
Photograph by courtesy National Portrait Gallery

JOHN LOCKE

AN ESSAY CONCERNING HUMAN UNDERSTANDING

EDITED WITH
AN INTRODUCTION, CRITICAL APPARATUS
AND GLOSSARY BY
PETER H. NIDDITCH

OXFORD
AT THE CLARENDON PRESS
1975

BOOK III

CHAPTER I

Of Words or Language in General.

§ 1. GOD having designed Man for a sociable Creature, made him not only with an inclination, and under a necessity to have fellowship with those of his own kind; but furnished him also with Language, which was to be the great Instrument, and common Tye of Society. *Man* therefore had by Nature his Organs so fashioned, as to be *fit to frame articulate Sounds*, which we call Words. But this was not enough to produce Language; for Parrots, and several other Birds, will be taught to make articulate Sounds distinct enough, which yet, by no means, are capable of Language.

§ 2. Besides articulate Sounds therefore, it was farther necessary, that he should be *able to use these Sounds, as Signs of internal Conceptions*; and to make them stand as marks for the *Ideas* within his own Mind, whereby they might be made known to others, and the Thoughts of Men's Minds be conveyed from one to another.

§ 3. But neither was this sufficient to make Words so useful as they ought to be. It is not enough for the perfection of Language, that Sounds can be made signs of *Ideas*, unless those *signs* can be so made use of, as *to comprehend several particular Things*: For the multiplication of Words would have perplexed their Use, had every particular thing need of a distinct name to be signified by. To remedy this inconvenience, Language had yet a farther improvement in the use of general Terms, whereby one word was made to mark a multitude of particular existences: Which advantageous use of Sounds was obtain'd only by the difference of the *Ideas* they were made signs of. Those names becoming general, which are made to stand for general *Ideas*, and those remaining particular, where the *Ideas* they are used for are particular.

§ 1. *Man fitted to form articulate Sounds.* § 2. *To make them signs of* Ideas. §§ 3, 4. *To make general Signs.*

(20–7) To . . . particular.] *add.* **2–5**

§ 4. Besides these Names which stand for *Ideas*, there be other words which Men make use of, not to signify any *Idea*, but the want or absence of some *Ideas* simple or complex, or all *Ideas* together; such as are *Nihil* in Latin, and in English, *Ignorance* and *Barrenness*. All which negative or privative Words, cannot be said properly to belong to, or signify no *Ideas*: for then they would be perfectly insignificant Sounds; but they relate to positive *Ideas*, and signify their absence.

§ 5. It may also lead us a little towards the Original of all our Notions and Knowledge, if we remark, how great a dependance our *Words* have on common sensible *Ideas*; and how those, which are made use of to stand for Actions and Notions quite removed from sense, *have their rise from thence, and from obvious sensible* Ideas *are transferred to more abstruse significations*, and made to stand for *Ideas* that come not under the cognizance of our senses; *v.g.* to *Imagine*, *Apprehend*, *Comprehend*, *Adhere*, *Conceive*, *Instill*, *Disgust*, *Disturbance*, *Tranquillity*, etc. are all Words taken from the Operations of sensible Things, and applied to certain Modes of Thinking. *Spirit*, in its primary signification, is Breath; *Angel*, a Messenger: And I doubt not, but if we could trace them to their sources, we should find, in all Languages, the names, which stand for Things that fall not under our Senses, to have had their first rise from sensible *Ideas*. By which we may give some kind of guess, what kind of Notions they were, and whence derived, which filled their Minds, who were the first Beginners of Languages; and now Nature, even in the naming of Things, unawares suggested to Men the Originals and Principles of all their Knowledge: whilst, to give Names, that might make known to others any Operations they felt in themselves, or any other *Ideas*, that came not under their Senses, they were fain to borrow Words from ordinary known *Ideas* of Sensation, by that means to make others the more easily to conceive those Operations they experimented in themselves, which made no outward sensible appearances; and then when they had got known and agreed Names, to signify those internal Operations of their own Minds,

§ 5. *Words ultimately derived from such as signify sensible* Ideas.

(1) Besides] **2–5** | *Words* then are made to be signs of our *Ideas*, and *are general or particular, as the Ideas they stand for are general or particular.* But besides **1** (1–2) other words] **2–5** | others **1** (2) Men] **2–5** | Men have found and **1** (4) *Nihil* in Latin] **2–5** | the Latin words, *Nihil* **1** (13–15) *rise* . . . senses] **2–5** | *Original*, and are transferred *from obvious sensible Ideas* **1** (20) sources] **2–5** | Originals **1**

they were sufficiently furnished to make known by Words, all their other *Ideas*; since they could consist of nothing, but either of outward sensible Perceptions, or of the inward Operations of their Minds about them; we having, as has been proved, no *Ideas* at all, but what originally come either from sensible Objects without, or what we feel within our selves, from the inward Workings of our own Spirits, which we are conscious to our selves of within.

§ 6. But to understand better the use and force of Language, as subservient to Instruction and Knowledge, it will be convenient to consider,

First, To what it is that Names, in the use of Language, are immediately applied.

Secondly, Since all (except proper) Names are general, and so stand not particularly for this or that single Thing; but for sorts and ranks of Things, it will be necessary to consider, in the next place, what the Sorts and Kinds, or, if you rather like the Latin Names, *what the Species and Genera of Things* are; wherein they consist; and how they come to be made. These being (as they ought) well looked into, we shall the better come to find the right use of Words; the natural Advantages and Defects of Language; and the remedies that ought to be used, to avoid the inconveniencies of obscurity or uncertainty in the signification of Words, without which, it is impossible to discourse with any clearness, or order, concerning Knowledge: Which being conversant about Propositions, and those most commonly universal ones, has greater connexion with Words, than perhaps is suspected.

These Considerations therefore, shall be the matter of the following Chapters.

CHAPTER II

Of the Signification of Words.

§ 1. MAN, though he have great variety of Thoughts, and such, from which others, as well as himself, might receive Profit and

§ 6. *Distribution.*
§ 1. *Words are sensible Signs necessary for Communication.*

(17–18) are; . . . consist;] **4er–5** | are, . . . consist, **1–4** (22) Words,] **4–5** | Words: **1–3** (*l. below* 30) **Coste** 'Division générale de ce Troisiéme Livre.'

Delight; yet they are all within his own Breast, invisible, and hidden from others, nor can of themselves be made appear. The Comfort, and Advantage of Society, not being to be had without Communication of Thoughts, it was necessary, that Man should find out some external sensible Signs, whereby those invisible *Ideas*, which his thoughts are made up of, might be made known to others. For this purpose, nothing was so fit, either for Plenty or Quickness, as those articulate Sounds, which with so much Ease and Variety, he found himself able to make. Thus we may conceive how *Words*, which were by Nature so well adapted to that purpose, come to be made use of by Men, as *the Signs of* their *Ideas*; not by any natural connexion, that there is between particular articulate Sounds and certain *Ideas*, for then there would be but one Language amongst all Men; but by a voluntary Imposition, whereby such a Word is made arbitrarily the Mark of such an *Idea*. The use then of Words, is to be sensible Marks of *Ideas*; and the *Ideas* they stand for, are their proper and immediate Signification.

§ 2. The use Men have of these Marks, being either to record their own Thoughts for the Assistance of their own Memory; or as it were, to bring out their *Ideas*, and lay them before the view of others: *Words in their primary or immediate Signification, stand for nothing, but the* Ideas *in the Mind of him that uses them*, how imperfectly soever, or carelesly those *Ideas* are collected from the Things, which they are supposed to represent. When a Man speaks to another, it is, that he may be understood; and the end of Speech is, that those Sounds, as Marks, may make known his *Ideas* to the Hearer. That then which Words are the Marks of, are the *Ideas* of the Speaker: Nor can any one apply them, as Marks, immediately to any thing else, but the *Ideas*, that he himself hath: For this would be to make them Signs of his own Conceptions, and yet apply them to other *Ideas*; which would be to make them Signs, and not Signs of his *Ideas* at the same time; and so in effect, to have no Signification at all. Words being voluntary Signs, they cannot be voluntary Signs imposed by him on Things he knows not. That would be

§§ 2, 3. *Words are the sensible Signs of his* Ideas *who uses them.*

(2) Comfort] **4–5** | Comfort therefore **1–3** (*likewise* **Coste**) (6) his . . . of] **4–5** | possess his Mind in so great variety **1–3** (7) this] **4–5** | which **1–3** (13) Sounds] **4–5** | Sounds, **1–3** (19) Thoughts] **4–5** | *Ideas* **1–3** (20) out their *Ideas*] **4–5** | them out **1–3** (21) others:] **4er–5** | others. **1–4** *or*] **5** | *and* **1–4** (25) Speech] **1T.er, 2–5** | the Speech **1** (30) Conceptions] **1er–5** | Conception **1**

to make them Signs of nothing, Sounds without Signification. A Man cannot make his Words the Signs either of Qualities in Things, or of Conceptions in the Mind of another, whereof he has none in his own. Till he has some *Ideas* of his own, he cannot suppose them to correspond with the Conceptions of another Man; nor can he use any Signs for them: For thus they would be the Signs of he knows not what, which is in Truth to be the Signs of nothing. But when he represents to himself other Men's *Ideas*, by some of his own, if he consent to give them the same Names, that other Men do, 'tis still to his own *Ideas*; to *Ideas* that he has, and not to *Ideas* that he has not.

§ 3. This is so necessary in the use of Language, that in this respect, the Knowing, and the Ignorant; the Learned, and Unlearned, use the *Words* they speak (with any meaning) all alike. They, *in every Man's Mouth, stand for the* Ideas *he has*, and which he would express by them. A Child having taken notice of nothing in the Metal he hears called Gold, but the bright shining yellow colour, he applies the Word Gold only to his own *Idea* of that Colour, and nothing else; and therefore calls the same Colour in a Peacocks Tail, Gold. Another that hath better observed, adds to shining yellow, great Weight: And then the Sound Gold, when he uses it, stands for a complex *Idea* of a shining Yellow and very weighty Substance. Another adds to those Qualities, Fusibility: and then the Word Gold to him signifies a Body, bright, yellow, fusible, and very heavy. Another adds Malleability. Each of these uses equally the Word Gold, when they have Occasion to express the *Idea*, which they have apply'd it to: But it is evident, that each can apply it only to his own *Idea*; nor can he make it stand, as a Sign of such a complex *Idea*, as he has not.

§ 4. But though Words, as they are used by Men, can properly and immediately signify nothing but the *Ideas*, that are in the Mind of the Speaker; yet they in their Thoughts give them a secret reference to two other things.

First, they suppose their Words to be Marks of the Ideas *in the Minds also of other Men, with whom they communicate*: For else they should talk in vain, and could not be understood, if the Sounds they applied to one *Idea*, were such, as by the Hearer, were applied to another,

§ 4. *Words often secretly referred, First, to the* Ideas *in other Mens Minds.*

(6) thus they] **4–5** | it **1–3** (7) Signs] **2–5** | Sign **1** (26) which] *add.* **4–5** (30–1) [*2nd*] the . . . Speaker] **2–5** | their Minds **1** (*l. below* 36) *In* **Coste**, §§ *4–6 come under the same marginal summary as that for* §§ *2, 3.* *other*] **2–3, 5** | *others* **4**

which is to speak two Languages. But in this, Men stand not usually to examine, whether the *Idea* they, and those they discourse with have in their Minds, be the same: But think it enough, that they use the Word, as they imagine, in the common Acceptation of that Language; in which case they suppose, that the *Idea*, they make it a Sign of, is precisely the same, to which the Understanding Men of that Country apply that Name.

§ 5. *Secondly*, Because *Men* would not be thought to talk *barely* of their own Imaginations, but of Things as really they are; therefore they *often suppose their Words to stand also for the reality of Things*. But this relating more particularly to Substances, and their Names, as perhaps the former does to simple *Ideas* and Modes, we shall speak of these two different ways of applying Words more at large, when we come to treat of the Names of mixed Modes, and Substances, in particular: Though give me leave here to say, that it is a perverting the use of Words, and brings unavoidable Obscurity and Confusion into their Signification, whenever we make them stand for any thing, but those *Ideas* we have in our own Minds.

§ 6. Concerning Words also it is farther to be considered. *First*, That they being immediately the Signs of Mens *Ideas*; and, by that means, the Instruments whereby Men communicate their Conceptions, and express to one another those Thoughts and Imaginations, they have within their own Breasts, *there comes by constant use*, to be such *a Connexion between certain Sounds, and the* Ideas *they stand for*, that the Names heard, almost as readily excite certain *Ideas*, as if the Objects themselves, which are apt to produce them, did actually affect the Senses. Which is manifestly so in all obvious sensible Qualities; and in all Substances, that frequently, and familiarly occur to us.

§ 7. *Secondly*, That though the proper and immediate Signification of Words, are *Ideas* in the Mind of the Speaker; yet because by familiar use from our Cradles, we come to learn certain articulate Sounds very perfectly, and have them readily on our Tongues, and always at hand in our Memories; but yet are not always careful to examine, or settle their Significations perfectly, it *often* happens that *Men*, even when they would apply themselves to an attentive

§ 5. *Secondly, To the reality of Things.* § 6. *Words by use readily excite* Ideas. § 7. *Words often used without signification.*

(2) those] **2–5** | he **1** (3) have in their Minds] *add.* **2–5** (5) case] **1–4**; *om.* **5** (19) it] **2–5** | this **1** (34) always . . . in] *add.* **2–5**

Consideration, do *set their Thoughts more on Words than Things*. Nay, because Words are many of them learn'd, before the *Ideas* are known for which they stand: Therefore some, not only Children, but Men, speak several Words, no otherwise than Parrots do, only because they have learn'd them, and have been accustomed to those Sounds. But so far as Words are of Use and Signification, so far is there a constant connexion between the Sound and the *Idea*; and a Designation, that the one stand for the other: without which Application of them, they are nothing but so much insignificant Noise.

§ 8. *Words* by long and familiar use, as has been said, come to excite in Men certain *Ideas*, so constantly and readily, that they are apt to suppose a natural connexion between them. But that they *signify* only Men's peculiar *Ideas*, and that *by a perfectly arbitrary Imposition*, is evident, in that they often fail to excite in others (even that use the same Language) the same *Ideas*, we take them to be the Signs of: And every Man has so inviolable a Liberty, to make Words stand for what *Ideas* he pleases, that no one hath the Power to make others have the same *Ideas* in their Minds, that he has, when they use the same Words, that he does. And therefore the great *Augustus* himself, in the Possession of that Power which ruled the World, acknowledged, he could not make a new Latin Word: which was as much as to say, that he could not arbitrarily appoint, what *Idea* any Sound should be a Sign of, in the Mouths and common Language of his Subjects. 'Tis true, common use, by a tacit Consent, appropriates certain Sounds to certain *Ideas* in all Languages, which so far limits the signification of that Sound, that unless a Man applies it to the same *Idea*, he does not speak properly: And let me add, that unless a Man's Words excite the same *Ideas* in the Hearer, which he makes them stand for in speaking, he does not speak intelligibly. But whatever be the consequence of any Man's using of Words differently, either from their general Meaning, or the particular Sense of the Person to whom he addresses them, this is certain, their signification, in his use of them, is limited to his *Ideas*, and they can be Signs of nothing else.

§§ 8–11. *Their Signification perfectly arbitrary.*

(27) does not] **4–5** | cannot **1–3** (28) let me add] **4–5** | it is also true **1–3**
(29) does not] **4–5** | cannot **1–3** (30–2) consequence . . . Person] **4–5** | consequence of any Man's use of Words different either from their Publick use, or that of the Persons **2–3** | consequences of his use of any Words, different either from the Publick, or that Person **1** (**1T.er** 'persons')

7 *On Saying That*

'I wish I had said that', said Oscar Wilde in applauding one of Whistler's witticisms. Whistler, who took a dim view of Wilde's originality, retorted, 'You will, Oscar; you will'.[1] This tale reminds us that an expression like 'Whistler said that' may on occasion serve as a grammatically complete sentence. Here we have, I suggest, the key to a correct analysis of indirect discourse, an analysis that opens a lead to an analysis of psychological sentences generally (sentences about propositional attitudes, so-called), and even, though this looks beyond anything to be discussed in the present paper, a clue to what distinguishes psychological concepts from others.

But let us begin with sentences usually deemed more representative of *oratio obliqua*, for example 'Galileo said that the earth moves' or 'Scott said that Venus is an inferior planet'. One trouble with such sentences is that we do not know their logical form. And to admit this is to admit that, whatever else we may know about them, we do not know the first thing. If we accept surface grammar as guide to logical form, we will see 'Galileo said that the earth moves' as containing the sentence 'the earth moves', and this sentence in turn as consisting of the singular term 'the earth', and a predicate, 'moves'. But if 'the earth' is, in this context, a singular term, it can be replaced, so far as the truth or falsity of the containing sentence is concerned, by any other singular term that refers to the same thing. Yet what seem like appropriate replacements can alter the truth of the original sentence.

The notorious apparent invalidity of this move can only be apparent, for the rule on which it is based no more than spells out

[1] From H. Jackson, *The Eighteen-Nineties*, 73.

DAVIDSON, DONALD
IN
INQUIRIES INTO
TRUTH AND MEANING

what is involved in the idea of a (logically) singular term. Only two lines of explanation, then, are open: we are wrong about the logical form, or we are wrong about the reference of the singular term.

What seems anomalous behaviour on the part of what seem singular terms dramatizes the problem of giving an orderly account of indirect discourse, but the problem is more pervasive. For what touches singular terms touches what they touch, and that is everything: quantifiers, variables, predicates, connectives. Singular terms refer, or pretend to refer, to the entities over which the variables of quantification range, and it is these entities of which the predicates are or are not true. So it should not surprise us that if we can make trouble for the sentence 'Scott said that Venus is an inferior planet' by substituting 'the Evening Star' for 'Venus', we can equally make trouble by substituting 'is identical with Venus or with Mercury' for the coextensive 'is an inferior planet'. The difficulties with indirect discourse cannot be solved simply by abolishing singular terms.

What should we ask of an adequate account of the logical form of a sentence? Above all, I would say, such an account must lead us to see the semantic character of the sentence—its truth or falsity—as owed to how it is composed, by a finite number of applications of some of a finite number of devices that suffice for the language as a whole, out of elements drawn from a finite stock (the vocabulary) that suffices for the language as a whole. To see a sentence in this light is to see it in the light of a theory for its language, a theory that gives the form of every sentence in that language. A way to provide such a theory is by recursively characterizing a truth predicate, along the lines suggested by Tarski.[2]

Two closely linked considerations support the idea that the structure with which a sentence is endowed by a theory of truth in Tarski's style deserves to be called the logical form of the sentence. By giving such a theory, we demonstrate in a persuasive way that the language, though it consists in an indefinitely large number of sentences, can be comprehended by a creature with finite powers. A theory of truth may be said to supply an effective explanation of the semantic role of each significant expression in any of its appearances. Armed with the theory, we can always answer the question, 'What are these familiar words doing here?' by saying how they

[2] A. Tarski, 'The Concept of Truth in Formalized Languages'. See Essays 2, 4, and 5.

contribute to the truth conditions of the sentence. (This is not to assign a 'meaning', much less a reference, to every significant expression.)

The study of the logical form of sentences is often seen in the light of another interest, that of expediting inference. From this point of view, to give the logical form of a sentence is to catalogue the features relevant to its place on the logical scene, the features that determine what sentences it is a logical consequence of, and what sentences it has as logical consequences. A canonical notation graphically encodes the relevant information, making theory of inference simple, and practice mechanical where possible.

Obviously the two approaches to logical form cannot yield wholly independent results, for logical consequence is defined in terms of truth. To say a second sentence is a logical consequence of a first is to say, roughly, that the second is true if the first is no matter how the non-logical constants are interpreted. Since what we count as a logical constant can vary independently of the set of truths, it is clear that the two versions of logical form, though related, need not be identical. The relation, in brief, seems this. Any theory of truth that satisfies Tarski's criteria must take account of all truth-affecting iterative devices in the language. In the familiar languages for which we know how to define truth the basic iterative devices are reducible to the sentential connectives, the apparatus of quantification, and the description operator if it is primitive. Where one sentence is a logical consequence of another on the basis of quantificational structure alone, a theory of truth will therefore entail that if the first sentence is true, the second is. There is no point, then, in not including the expressions that determine quantificational structure among the logical constants, for when we have characterized truth, on which any account of logical consequence depends, we have already committed ourselves to all that calling such expressions logical constants could commit us. Adding to this list of logical constants will increase the inventory of logical truths and consequence-relations beyond anything a truth definition demands, and will therefore yield richer versions of logical form. For the purposes of the present paper, however, we can cleave to the most austere interpretations of logical consequence and logical form, those that are forced on us when we give a theory of truth.[3]

[3] For further defence of a concept of logical form based on a theory of truth, see *Essays on Actions and Events*, 137–46.

We are now in a position to explain our aporia over indirect discourse: what happens is that the relation between truth and consequence just sketched appears to break down. In a sentence like 'Galileo said that the earth moves' the eye and mind perceive familiar structure in the words 'the earth moves'. And structure there must be if we are to have a theory of truth at all, for an infinite number of sentences (all sentences in the indicative, apart from some trouble over tense) yield sense when plugged into the slot in 'Galileo said that _____'. So if we are to give conditions of truth for all the sentences so generated, we cannot do it sentence by sentence, but only by discovering an articulate structure that permits us to treat each sentence as composed of a finite number of devices that make a stated contribution to its truth conditions. As soon as we assign familiar structure, however, we must allow the consequences of that assignment to flow, and these, as we know, are in the case of indirect discourse consequences we refuse to buy. In a way, the matter is even stranger than that. Not only do familiar consequences fail to flow from what looks to be familiar structure, but our common sense of language feels little assurance in any inferences based on the words that follow the 'said that' of indirect discourse (there are exceptions).

So the paradox is this: on the one hand, intuition suggests, and theory demands, that we discover semantically significant structure in the 'content-sentences' of indirect discourse (as I shall call sentences following 'said that'). On the other hand, the failure of consequence-relations invites us to treat contained sentences as semantically inert. Yet logical form and consequence relations cannot be divorced in this way.

One proposal at this point is to view the words that succeed the 'said that' as operating within concealed quotation marks, their sole function being to help refer to a sentence, and their semantic inertness explained by an account of quotation. One drawback of this proposal is that no usual account of quotation is acceptable; even by the minimal standards we have set for an account of logical form. For according to most stories, quotations are singular terms without significant semantic structure, and since there must be an infinite number of different quotations, no language that contains them can have a recursively defined truth predicate. This may be taken to show that the received accounts of quotation must be mistaken—I think it does. But then we can hardly pretend that we

have solved the problem of indirect discourse by appeal to quotation.[4]

Perhaps it is not hard to invent a theory of quotation that will serve: the following theory is all but explicit in Quine. Simply view quotations as abbreviations for what you get if you follow these instructions: to the right of the first letter that has opening quotation marks on its left write right-hand quotation marks, then the sign for concatenation, and then left-hand quotation marks, in that order; do this after each letter (treating punctuation signs as letters) until you reach the terminating right-hand quotation marks. What you now have is a complex singular term that gives what Tarski calls a structural description of an expression. There is a modest addition to vocabulary: names of letters and of punctuation signs, and the sign for concatenation. There is a corresponding addition to ontology: letters and punctuation signs. And finally, if we carry out the application to sentences in indirect discourse, there will be the logical consequences that the new structure dictates. For two examples, each of the following will be entailed by 'Galileo said that the earth moves':

$(\exists x)$ (Galileo said that 'the ea'$^\frown x^\frown$'th moves')

and (with the premise 'r = the 18th letter in the alphabet'):

Galileo said that 'the ea'$^\frown$the 18th letter of the alphabet$^\frown$'th moves'

(I have clung to abbreviations as far as possible.) These inferences are not meant in themselves as criticism of the theory of quotation; they merely illuminate it.

Quine discusses the quotational approach to indirect discourse in *Word and Object*,[5] and abandons it for what seems, to me, a wrong reason. Not that there is not a good reason; but to appreciate *it* is to be next door to a solution, as I shall try to show.

Let us follow Quine through the steps that lead him to reject the quotational approach. The version of the theory he considers is not the one once proposed by Carnap to the effect that 'said that' is a two-place predicate true of ordered pairs of people and sentences.[6]

[4] See Essays 1 and 6.

[5] W. V. Quine, *Word and Object*, Ch. 6. Hereafter numerals in parentheses refer to pages of this book.

[6] R. Carnap, *The Logical Syntax of Language*, 248. The same was in effect proposed by P. T. Geach in *Mental Acts*.

The trouble with this idea is not that it forces us to assimilate indirect discourse to direct, for it does not. The 'said that' of indirect discourse, like the 'said' of direct, may relate persons and sentences, but be a different relation; the former, unlike the latter, may be true of a person, and a sentence he never spoke in a language he never knew. The trouble lies rather in the chance that the same sentence may have different meanings in different languages—not too long a chance either if we count ideolects as languages. To give an example, the sounds 'Empedokles liebt' do fairly well as a German or an English sentence, in one case saying that Empedokles loved and in the other telling us what he did from the top of Etna. If we analyse 'Galileo said that the earth moves' as asserting a relation between Galileo and the sentence 'The earth moves', we do not have to assume that Galileo spoke English, but we cannot avoid the assumption that the words of the content-sentence are to be understood as an English sentence.[7]

Calling the relativity to English an assumption may be misleading; perhaps the reference to English is explicit, as follows. A long-winded version of our favourite sentence might be 'Galileo spoke a sentence that meant in his language what "The earth moves" means in English'. Since in this version it needs all the words except 'Galileo' and 'The earth moves' to do the work of 'said that', we must count the reference to English as explicit in the 'said that'. To see how odd this is, however, it is only necessary to reflect that the English words 'said that', with their built-in reference to English, would no longer translate (by even the roughest extensional standards) the French 'dit que'.

We can shift the difficulty over translation away from the 'said that' or 'dit que' by taking these expressions as three-place predicates relating a speaker, a sentence, and a language, the reference to a language to be supplied either by our (in practice nearly infallible) knowledge of the language to which the quoted material is to be taken as belonging, or by a demonstrative reference to the language of the entire sentence. Each of these suggestions has its own appeal, but neither leads to an analysis that will pass the translation test. To take the demonstrative proposal, translation into French will carry 'said that' into 'dit que', the demonstrative reference will automatically, and hence perhaps still within the

[7] The point is due to A. Church, 'On Carnap's Analysis of Statements of Assertion and Belief'.

bounds of strict translation, shift from English to French. But when we translate the final singular term, which names an English sentence, we produce a palpably false result.

These exercises help bring out important features of the quotational approach. But now it is time to remark that there would be an anomaly in a position, like the one under consideration, that abjured reference to propositions in favour of reference to languages. For languages (as Quine remarks in a similar context in *Word and Object*) are at least as badly individuated, and for much the same reasons, as propositions. Indeed, an obvious proposal linking them is this: languages are identical when identical sentences express identical propositions. We see, then, that quotational theories of indirect discourse, those we have discussed anyway, cannot claim an advantage over theories that frankly introduce intensional entities from the start; so let us briefly consider theories of the latter sort.

It might be thought, and perhaps often is, that if we are willing to welcome intensional entities without stint—properties, propositions, individual concepts, and whatever else—then no further difficulties stand in the way of giving an account of the logical form of sentences in *oratio obliqua*. This is not so. Neither the languages Frege suggests as models for natural languages nor the languages described by Church are amenable to theory in the sense of a truth definition meeting Tarski's standards.[8] What stands in the way in Frege's case is that every referring expression has an infinite number of entities it may refer to, depending on the context, and there is no rule that gives the reference in more complex contexts on the basis of the reference in simpler ones. In Church's languages, there is an infinite number of primitive expressions; this directly blocks the possibility of recursively characterizing a truth predicate satisfying Tarski's requirements.

Things might be patched up by following a leading idea of Carnap's *Meaning and Necessity* and limiting the semantic levels to two: extensions and (first-level) intensions.[9] An attractive strategy might then be to turn Frege, thus simplified, upside down by letting each singular term refer to its sense or intension, and providing a

[8] G. Frege, 'On Sense and Reference'; A. Church, 'A Formulation of the Logic of Sense and Denotation'. See Essay 1.

[9] The idea of an essentially Fregean approach limited to two semantic levels has been suggested by M. Dummett in *Frege: Philosophy of Language*, Ch. 9.

reality function (similar to Church's delta function) to map intensions on to extensions. Under such treatment our sample sentence would emerge like this: 'The reality of Galileo said that the earth moves.' Here we must suppose that 'the earth' names an individual concept which the function referred to by 'moves' maps on to the proposition that the earth moves; the function referred to by 'said that' in turn maps Galileo and the proposition that the earth moves on to a truth value. Finally, the name 'Galileo' refers to an individual concept which is mapped, by the function referred to by 'the reality of' on to Galileo. With ingenuity, this theory can perhaps be made to accommodate quantifiers that bind variables both inside and outside contexts created by verbs like 'said' and 'believes'. There is no special problem about defining truth for such a language: everything is on the up and up, purely extensional save in ontology. This seems to be a theory that might do all we have asked. Apart from nominalistic qualms, why not accept it?

My reasons against this course are essentially Quine's. Finding right words of my own to communicate another's saying is a problem in translation (216–17). The words I use in the particular case may be viewed as products of my total theory (however vague and subject to correction) of what the originating speaker means by anything he says: such a theory is indistinguishable from a characterization of a truth predicate, with his language as object language and mine as metalanguage. The crucial point is that there will be equally acceptable alternative theories which differ in assigning clearly non-synonymous sentences of mine as translations of his same utterance. This is Quine's thesis of the indeterminacy of translation (218–21).[10] An example will help bring out the fact that the thesis applies not only to translation between speakers of conspicuously different languages, but also to cases nearer home.

Let someone say (and now discourse is direct), 'There's a hippopotamus in the refrigerator'; am I necessarily right in reporting him as having said that there is a hippopotamus in the refrigerator? Perhaps; but under questioning he goes on, 'It's roundish, has a wrinkled skin, does not mind being touched. It has a pleasant taste, at least the juice, and it costs a dime. I squeeze two or three for breakfast.' After some finite amount of such talk we slip over the line where it is plausible or even possible to say correctly

[10] My assimilation of a translation manual to a theory of truth is not in Quine. For more on this and related matters, see Essays 2, 11, and 16.

that he said there was a hippopotamus in the refrigerator, for it becomes clear he means something else by at least some of his words than I do. The simplest hypothesis so far is that my word 'hippopotamus' no longer translates his word 'hippopotamus'; my word 'orange' might do better. But in any case, long before we reach the point where homophonic translation must be abandoned, charity invites departures. Hesitation over whether to translate a saying of another by one or another of various non-synonymous sentences of mine does not necessarily reflect a lack of information: it is just that beyond a point there is no deciding, even in principle, between the view that the Other has used words as we do but has more or less weird beliefs, and the view that we have translated him wrong. Torn between the need to make sense of a speaker's words and the need to make sense of the pattern of his beliefs, the best we can do is choose a theory of translation that maximizes agreement. Surely there is no future in supposing that in earnestly uttering the words 'There's a hippopotamus in the refrigerator' the Other has disagreed with us about what can be in the refrigerator if we also must then find ourselves disagreeing with him about the size, shape, colour, manufacturer, horsepower, and wheelbase of hippopotami.

None of this shows there is no such thing as correctly reporting, through indirect discourse, what another has said. All that the indeterminacy shows is that if there is one way of getting it right there are other ways that differ substantially in that non-synonymous sentences are used after 'said that'. And this is enough to justify our feeling that there is something bogus about the sharpness questions of meaning must in principle have if meanings are entities.

The lesson was implicit in a discussion started some years ago by Benson Mates. Mates claimed that the sentence 'Nobody doubts that whoever believes that the seventh consulate of Marius lasted less than a fortnight believes that the seventh consulate of Marius lasted less than a fortnight' is true and yet might well become false if the last word were replaced by the (supposed synonymous) words 'period of fourteen days', and that this could happen no matter what standards of synonomy we adopt short of the question-begging 'substitutable everywhere *salva veritate*'.[11] Church and Sellars responded by saying the difficulty could be resolved by firmly distinguishing between substitutions based on the speaker's use of

[11] B. Mates, 'Synonymity'. The example is Church's.

language and substitutions coloured by the use attributed to others.[12] But this is a solution only if we think there is some way of telling, in what another says, what is owed to the meanings he gives his words and what to his beliefs about the world. According to Quine, this is a distinction that cannot be drawn.

The detour has been lengthy; I return now to Quine's discussion of the quotational approach in *Word and Object*. As reported above, Quine rejects relativization to a language on the grounds that the principle of the individuation of languages is obscure, and the issue when languages are identical irrelevant to indirect discourse (214). He now suggests that instead of interpreting the content-sentence of indirect discourse as occurring in a language, we interpret it as voiced by a speaker at a time. The speaker and time relative to which the content-sentence needs understanding is, of course, the speaker of that sentence, who is thereby indirectly attributing a saying to another. So now 'Galileo said that the earth moves' comes to mean something like 'Galileo spoke a sentence that in his mouth meant what "The earth moves" now means in mine'. Quine makes no objection to this proposal because he thinks he has something simpler and at least as good in reserve. But in my opinion the present proposal deserves more serious consideration, for I think it is nearly right, while Quine's preferred alternatives are seriously defective.

The first of these alternatives is Scheffler's inscriptional theory.[13] Scheffler suggests that sentences in indirect discourse relate a speaker and an utterance: the role of the content-sentence is to help convey what sort of utterance it was. What we get this way is, 'Galileo spoke a that-the-earth-moves utterance'. The predicate '*x* is-a-that-the-earth-moves-utterance' has, so far as theory of truth and of inference are concerned, the form of an unstructured one-place predicate. Quine does not put the matter quite this way, and he may resist my appropriation of the terms 'logical form' and 'structure' for purposes that exclude application to Scheffler's predicate. Quine calls the predicate 'compound' and describes it as composed of an operator and a sentence (214, 215). These are matters of terminology; the substance, about which there may be no disagreement, is that on Scheffler's theory sentences in *oratio obliqua* have no logical relations that depend on structure in the predicate,

[12] A. Church, 'Intensional Isomorphism and Identity of Belief'; W. Sellars, 'Putnam on Synonymity and Belief'.

[13] I. Scheffler, 'An Inscriptional Approach to Indirect Quotation'.

and a truth predicate that applies to all such sentences cannot be characterized in Tarski's style. The reason is plain: there is an infinite number of predicates with the syntax '*x* is-a-_____-utterance' each of which is, in the eyes of semantic theory, unrelated to the rest.

Quine has seized one horn of the dilemma. Since attributing semantic structure to content-sentences in indirect discourse apparently forces us to endorse logical relations we do not want, Quine gives up the structure. The result is that another desideratum of theory is neglected, that truth be defined.

Consistent with his policy of renouncing structure that supports no inferences worth their keep, Quine contemplates one further step; he says, '... a final alternative that I find as appealing as any is simply to dispense with the objects of the propositional attitudes' (216). Where Scheffler still saw 'said that' as a two-place predicate relating speakers and utterances, though welding content-sentences into one-piece one-place predicates true of utterances, Quine now envisions content-sentence and 'said that' welded directly to form the one-place predicate '*x* said-that-the-earth-moves', true of persons. Of course some inferences inherent in Scheffler's scheme now fall away: we can no longer infer 'Galileo said something' from our sample sentence, nor can we infer from it and 'Someone denied that the earth moves' the sentence 'Someone denied what Galileo said'. Yet as Quine reminds us, inferences like these may fail on Scheffler's analysis too when the analysis is extended along the obvious line to belief and other propositional attitudes, since needed utterances may fail to materialize (215). The advantages of Scheffler's theory over Quine's 'final alternative' are therefore few and uncertain; this is why Quine concludes that the view that invites the fewest inferences is 'as appealing as any'.

This way of eliminating unwanted inferences unfortunately abolishes most of the structure needed by the theory of truth. So it is worth returning for another look at the earlier proposal to analyse indirect discourse in terms of a predicate relating an originating speaker, a sentence, and the present speaker of the sentence in indirect discourse. For that proposal did not cut off any of the simple entailments we have been discussing, and it alone of recent suggestions promised, when coupled with a workable theory of quotation, to yield to standard semantic methods. But there is a subtle flaw.

We tried to bring out the flavour of the analysis to which we have returned by rewording our favourite sentence as 'Galileo uttered a sentence that meant in his mouth what "The earth moves" means now in mine'. We should not think ill of this verbose version of 'Galileo said that the earth moves' because of apparent reference to a meaning ('what "The earth moves" means'); this expression is not treated as a singular term in the theory. We are indeed asked to make sense of a judgement of synonomy between utterances, but not as the foundation of a theory of language, merely as an unanalysed part of the content of the familiar idiom of indirect discourse. The idea that underlies our awkward paraphrase is that of *samesaying*: when I say that Galileo said that the earth moves, I represent Galileo and myself as samesayers.[14]

And now the flaw is this. If I merely *say* we are samesayers, Galileo and I, I have yet to *make* us so; and how am I to do this? Obviously, by saying what he said; not by using his words (necessarily), but by using words the same in import here and now as his then and there. Yet this is just what, on the theory, I cannot do. For the theory brings the content-sentence into the act sealed in quotation marks, and on any standard theory of quotation, this means the content-sentence is mentioned and not used. In uttering the words 'The earth moves' I do not, according to this account, say anything remotely like what Galileo is claimed to have said; I do not, in fact, say anything. My words in the frame provided by 'Galileo said that_____' merely help refer to a sentence. There will be no missing the point if we expand quotation in the style we recently considered. Any intimation that Galileo and I are samesayers vanishes in this version:

Galileo said that 'T'⌢'h'⌢'e'⌢ '⌢'e'⌢'a'⌢'r'⌢'t'⌢'h'⌢ '⌢'m'⌢'o'⌢'v'⌢'e'⌢'s'

[14] Strictly speaking, the verb 'said' is here analysed as a three-place predicate which holds of a speaker (Galileo), an utterance of the speaker ('Eppur si muove'), and an utterance of the attributer ('The earth moves'). This predicate is from a semantic point of view a primitive. The fact that an informal paraphrase of the predicate appeals to a relation of sameness of content as between utterances introduces no intentional entities or semantics. Some have regarded this as a form of cheating, but the policy is deliberate and principled. For a discussion of the distinction between questions of logical form (which is the present concern) and the analysis of individual predicates, see Essay 2. It is also worth observing that radical interpretation, if it succeeds, yields an adequate concept of synonymy as between utterances. See the end of Essay 12. [Footnote added in 1982.]

We seem to have been taken in by a notational accident, a way of referring to expressions that when abbreviated produces framed pictures of the very words referred to. The difficulty is odd; let's see if we can circumvent it. Imagine an altered case. Galileo utters his words 'Eppur si muove', I utter my words, 'The earth moves'. There is no problem yet in recognizing that we are samesayers; an utterance of mine matches an utterance of his in purport. I am not now using my words to help refer to a sentence; I speak for myself, and my words refer in their usual way to the earth and to its movement. If Galileo's utterance 'Eppur si muove' made us samesayers, then some utterance or other of Galileo's made us samesayers. The form '($\exists x$) (Galileo's utterance x and my utterance y makes us samesayers)' is thus a way of attributing any saying I please to Galileo provided I find a way of replacing 'y' by a word or phrase that refers to an appropriate utterance of mine. And surely there is a way I can do this: I need only produce the required utterance and replace 'y' by a reference to it. Here goes:

The earth moves.

($\exists x$) (Galileo's utterance x and my last utterance makes us samesayers).

Definitional abbreviation is all that is needed to bring this little skit down to:

The earth moves.
Galileo said that.

Here the 'that' is a demonstrative singular term referring to an utterance (not a sentence).

This form has a small drawback in that it leaves the hearer up in the air about the purpose served by saying 'The earth moves' until the act has been performed. As if, say, I were first to tell a story and then add, 'That's how it was once upon a time'. There's some fun to be had this way, and in any case no amount of telling what the illocutionary force of our utterances is is going to insure that they have that force. But in the present case nothing stands in the way of reversing the order of things, thus:

Galileo said that.
The earth moves.

It is now safe to allow a tiny orthographic change, a change without semantic significance, but suggesting to the eye the relation of introducer and introduced: we may suppress the stop after 'that' and the consequent capitalization:

Galileo said that the earth moves.

Perhaps it should come as no surprise to learn that the form of psychological sentences in English apparently evolved in much the way these ruminations suggest. According to the *Oxford English Dictionary*,

> The use of *that* is generally held to have arisen out of the demonstrative pronoun pointing to the clause which it introduces. Cf. (1) He once lived here: we all know *that*; (2) *That* (now *this*) we all know: he once lived here; (3) We all know *that* (or *this*): he once lived here; (4) We all know *that* he once lived here . . .[15]

The proposal then is this: sentences in indirect discourse, as it happens, wear their logical form on their sleeves (except for one small point). They consist of an expression referring to a speaker, the two-place predicate 'said', and a demonstrative referring to an utterance. Period. What follows gives the content of the subject's saying, but has no logical or semantic connection with the original attribution of a saying. This last point is no doubt the novel one, and upon it everything depends: from a semantic point of view the content-sentence in indirect discourse is not contained in the sentence whose truth counts, i.e. the sentence that ends with 'that'.

We would do better, in coping with this subject, to talk of inscriptions and utterances and speech acts, and avoid reference to sentences.[16] For what an utterance of 'Galileo said that' does is announce a further utterance. Like any utterance, this first may be serious or silly, assertive or playful; but if it is true, it must be followed by an utterance synonymous with some other. The second utterance, the introduced act, may also be true or false, done in the mode of assertion or of play. But if it is as announced, it must serve

[15] J. A. H. Murray *et al.* (eds.), *The Oxford English Dictionary*, 253. Cf. C. T. Onions, *An Advanced English Syntax*, 154–6. I first learned that 'that' in such contexts evolved from an explicit demonstrative in J. Hintikka, *Knowledge and Belief*, 13. Hintikka remarks that a similar development has taken place in German and Finnish. I owe the *OED* reference to Eric Stiezel.

[16] I assume that a theory of truth for a language containing demonstratives must apply strictly to utterances and not to sentences, or will treat truth as a relation between sentences, speakers, and times. See Essays 2 and 4.

at least the purpose of conveying the content of what someone said. The role of the introducing utterance is not unfamiliar: we do the same with words like 'This is a joke', 'This is an order', 'He commanded that', 'Now hear this'. Such expressions might be called performatives, for they are used to usher in performances on the part of the speaker. A certain interesting reflexive effect sets in when performatives occur in the first-person present tense, for then the speaker utters words which if true are made so exclusively by the content and mode of the performance that follows, and the mode of this performance may well be in part determined by that same performative introduction. Here is an example that will also provide the occasion for a final comment on indirect discourse.

'Jones asserted that Entebbe is equatorial' would, if we parallel the analysis of indirect discourse, come to mean something like, 'An utterance of Jones' in the assertive mode had the content of this utterance of mine. Entebbe is equatorial.' The analysis does not founder because the modes of utterance of the two speakers may differ; all that the truth of the performative requires is that the second utterance, in whatever mode (assertive or not), match in content an assertive utterance of Jones. Whether such an asymmetry is appropriate in indirect discourse depends on how much of assertion we read into the concept of saying. Now suppose I try: 'I assert that Entebbe is equatorial.' Of course by saying this I may not assert anything; mood of words cannot guarantee mode of utterance. But if my utterance of the performative is true, then do I say something in the assertive mode that has the content of my second utterance—I do, that is, assert that Entebbe is equatorial. If I do assert it, an element in my success is no doubt my utterance of the performative, which announces an assertion; thus performatives tend to be self-fulfilling. Perhaps it is this feature of performatives that has misled some philosophers into thinking that performatives, or their utterances, are neither true nor false.

On the analysis of indirect discourse here proposed, standard problems seem to find a just solution. The appearance of failure of the laws of extensional substitution is explained as due to our mistaking what are really two sentences for one: we make substitutions in one sentence, but it is the other (the utterance of) which changes in truth. Since an utterance of 'Galileo said that' and any utterance following it are semantically independent, there is no reason to predict, on grounds of form alone, any *particular* effect on

the truth of the first from change in the second. On the other hand, if the second utterance had been different in any way at all, the first utterance *might* have had a different truth value, for the reference of the 'that' would have changed.

The paradox, that sentences (utterances) in *oratio obliqua* do not have the logical consequences they should if truth is to be defined, is resolved. What follows the verb 'said' has only the structure of a singular term, usually the demonstrative 'that'. Assuming the 'that' refers, we can infer that Galileo said something from 'Galileo said that'; but this is welcome. The familiar words coming in the train of the performative of indirect discourse do, on my account, have structure, but it is familiar structure and poses no problem for theory of truth not there before indirect discourse was the theme.

Since Frege, philosophers have become hardened to the idea that content-sentences in talk about propositional attitudes may strangely refer to such entities as intensions, propositions, sentences, utterances, and inscriptions. What is strange is not the entities, which are all right in their place (if they have one), but the notion that ordinary words for planets, people, tables, and hippopotami in indirect discourse may give up these pedestrain references for the exotica. If we could recover our pre-Fregean semantic innocence, I think it would seem to us plainly incredible that the words 'The earth moves', uttered after the words 'Galileo said that', mean anything different, or refer to anything else, than is their wont when they come in other environments. No doubt their role in *oratio obliqua* is in some sense special; but that is another story. Language is the instrument it is because the same expression, with semantic features (meaning) unchanged, can serve countless purposes. I have tried to show how our understanding of indirect discourse does not strain this basic insight.

DAVIDSON, DONALD

IN:

INQUIRIES INTO TRUTH AND MEANING

2 *Truth and Meaning*

It is conceded by most philosophers of language, and recently by some linguists, that a satisfactory theory of meaning must give an account of how the meanings of sentences depend upon the meanings of words. Unless such an account could be supplied for a particular language, it is argued, there would be no explaining the fact that we can learn the language: no explaining the fact that, on mastering a finite vocabulary and a finitely stated set of rules, we are prepared to produce and to understand any of a potential infinitude of sentences. I do not dispute these vague claims, in which I sense more than a kernel of truth.[1] Instead I want to ask what it is for a theory to give an account of the kind adumbrated.

One proposal is to begin by assigning some entity as meaning to each word (or other significant syntactical feature) of the sentence; thus we might assign Theaetetus to 'Theaetetus' and the property of flying to 'flies' in the sentence 'Theaetetus flies'. The problem then arises how the meaning of the sentence is generated from these meanings. Viewing concatenation as a significant piece of syntax, we may assign to it the relation of participating in or instantiating; however, it is obvious that we have here the start of an infinite regress. Frege sought to avoid the regress by saying that the entities corresponding to predicates (for example) are 'unsaturated' or 'incomplete' in contrast to the entities that correspond to names, but this doctrine seems to label a difficulty rather than solve it.

The point will emerge if we think for a moment of complex singular terms, to which Frege's theory applies along with sentences. Consider the expression 'the father of Annette'; how does the

[1] See Essay 1.

meaning of the whole depend on the meaning of the parts? The answer would seem to be that the meaning of 'the father of' is such that when this expression is prefixed to a singular term the result refers to the father of the person to whom the singular term refers. What part is played, in this account, by the unsaturated or incomplete entity for which 'the father of' stands? All we can think to say is that this entity 'yields' or 'gives' the father of x as value when the argument is x, or perhaps that this entity maps people on to their fathers. It may not be clear whether the entity for which 'the father of' is said to stand performs any genuine explanatory function as long as we stick to individual expressions; so think instead of the infinite class of expressions formed by writing 'the father of' zero or more times in front of 'Annette'. It is easy to supply a theory that tells, for an arbitrary one of these singular terms, what it refers to: if the term is 'Annette' it refers to Annette, while if the term is complex, consisting of 'the father of' prefixed to a singular term t, then it refers to the father of the person to whom t refers. It is obvious that no entity corresponding to 'the father of' is, or needs to be, mentioned in stating this theory.

It would be inappropriate to complain that this little theory *uses* the words 'the father of' in giving the reference of expressions containing those words. For the task was to give the meaning of all expressions in a certain infinite set on the basis of the meaning of the parts; it was not in the bargain also to give the meanings of the atomic parts. On the other hand, it is now evident that a satisfactory theory of the meanings of complex expressions may not require entities as meanings of all the parts. It behoves us then to rephrase our demand on a satisfactory theory of meaning so as not to suggest that individual words must have meanings at all, in any sense that transcends the fact that they have a systematic effect on the meanings of the sentences in which they occur. Actually, for the case at hand we can do better still in stating the criterion of success: what we wanted, and what we got, is a theory that entails every sentence of the form 't refers to x' where 't' is replaced by a structural description[2] of a singular term, and 'x' is replaced by that term itself. Further, our theory accomplishes this without appeal to any semantical concepts beyond the basic 'refers to'. Finally, the theory

[2] A 'structural description' of an expression describes the expression as a concatenation of elements drawn from a fixed finite list (for example of words or letters).

clearly suggests an effective procedure for determining, for any singular term in its universe, what that term refers to.

A theory with such evident merits deserves wider application. The device proposed by Frege to this end has a brilliant simplicity: count predicates as a special case of functional expressions, and sentences as a special case of complex singular terms. Now, however, a difficulty looms if we want to continue in our present (implicit) course of identifying the meaning of a singular term with its reference. The difficulty follows upon making two reasonable assumptions: that logically equivalent singular terms have the same reference, and that a singular term does not change its reference if a contained singular term is replaced by another with the same reference. But now suppose that 'R' and 'S' abbreviate any two sentences alike in truth value. Then the following four sentences have the same reference:

(1) R
(2) $\hat{x}(x = x.R) = \hat{x}(x = x)$
(3) $\hat{x}(x = x.S) = \hat{x}(x = x)$
(4) S

For (1) and (2) are logically equivalent, as are (3) and (4), while (3) differs from (2) only in containing the singular term '$\hat{x}(x = x.S)$' where (2) contains '$\hat{x}(x = x.R)$' and these refer to the same thing if S and R are alike in truth value. Hence any two sentences have the same reference if they have the same truth value.[3] And if the meaning of a sentence is what it refers to, all sentences alike in truth value must be synonymous—an intolerable result.

Apparently we must abandon the present approach as leading to a theory of meaning. This is the natural point at which to turn for help to the distinction between meaning and reference. The trouble, we are told, is that questions of reference are, in general, settled by extra-linguistic facts, questions of meaning not, and the facts can conflate the references of expressions that are not synonymous. If we want a theory that gives the meaning (as distinct from reference) of each sentence, we must start with the meaning (as distinct from reference) of the parts.

Up to here we have been following in Frege's footsteps; thanks to

[3] The argument derives from Frege. See A. Church, *Introduction to Mathematical Logic*, 24–5. It is perhaps worth mentioning that the argument does not depend on any particular identification of the entities to which sentences are supposed to refer.

him, the path is well known and even well worn. But now, I would like to suggest, we have reached an impasse: the switch from reference to meaning leads to no useful account of how the meanings of sentences depend upon the meanings of the words (or other structural features) that compose them. Ask, for example, for the meaning of 'Theaetetus flies'. A Fregean answer might go something like this: given the meaning of 'Theaetetus' as argument, the meaning of 'flies' yields the meaning of 'Theaetetus flies' as value. The vacuity of this answer is obvious. We wanted to know what the meaning of 'Theaetetus flies' is; it is no progress to be told that it is the meaning of 'Theaetetus flies'. This much we knew before any theory was in sight. In the bogus account just given, talk of the structure of the sentence and of the meanings of words was idle, for it played no role in producing the given description of the meaning of the sentence.

The contrast here between a real and pretended account will be plainer still if we ask for a theory, analogous to the miniature theory of reference of singular terms just sketched, but different in dealing with meanings in place of references. What analogy demands is a theory that has as consequences all sentences of the form '*s* means *m*' where '*s*' is replaced by a structural description of a sentence and '*m*' is replaced by a singular term that refers to the meaning of that sentence; a theory, moreover, that provides an effective method for arriving at the meaning of an arbitrary sentence structurally described. Clearly some more articulate way of referring to meanings than any we have seen is essential if these criteria are to be met.[4] Meanings as entities, or the related concept of synonymy, allow us to formulate the following rule relating sentences and their parts: sentences are synonymous whose corresponding parts are synonymous ('corresponding' here needs spelling out of course). And meanings as entities may, in theories such as Frege's, do duty, on occasion, as references, thus losing their status as entities distinct from references. Paradoxically, the one thing meanings do not seem to do is oil the wheels of a theory of meaning—at least as long as we require of such a theory that it non-trivially give the meaning of

[4] It may be thought that Church, in 'A Formulation of the Logic of Sense and Denotation', has given a theory of meaning that makes essential use of meanings as entities. But this is not the case: Church's logics of sense and denotation are interpreted as being about meanings, but they do not mention expressions and so cannot of course be theories of meaning in the sense now under discussion.

every sentence in the language. My objection to meanings in the theory of meaning is not that they are abstract or that their identity conditions are obscure, but that they have no demonstrated use.

This is the place to scotch another hopeful thought. Suppose we have a satisfactory theory of syntax for our language, consisting of an effective method of telling, for an arbitrary expression, whether or not it is independently meaningful (i.e. a sentence), and assume as usual that this involves viewing each sentence as composed, in allowable ways, out of elements drawn from a fixed finite stock of atomic syntactical elements (roughly, words). The hopeful thought is that syntax, so conceived, will yield semantics when a dictionary giving the meaning of each syntactic atom is added. Hopes will be dashed, however, if semantics is to comprise a theory of meaning in our sense, for knowledge of the structural characteristics that make for meaningfulness in a sentence, plus knowledge of the meanings of the ultimate parts, does not add up to knowledge of what a sentence means. The point is easily illustrated by belief sentences. Their syntax is relatively unproblematic. Yet, adding a dictionary does not touch the standard semantic problem, which is that we cannot account for even as much as the truth conditions of such sentences on the basis of what we know of the meanings of the words in them. The situation is not radically altered by refining the dictionary to indicate which meaning or meanings an ambiguous expression bears in each of its possible contexts; the problem of belief sentences persists after ambiguities are resolved.

The fact that recursive syntax with dictionary added is not necessarily recursive semantics has been obscured in some recent writing on linguistics by the intrusion of semantic criteria into the discussion of purportedly syntactic theories. The matter would boil down to a harmless difference over terminology if the semantic criteria were clear; but they are not. While there is agreement that it is the central task of semantics to give the semantic interpretation (the meaning) of every sentence in the language, nowhere in the linguistic literature will one find, so far as I know, a straightforward account of how a theory performs this task, or how to tell when it has been accomplished. The contrast with syntax is striking. The main job of a modest syntax is to characterize *meaningfulness* (or sentencehood). We may have as much confidence in the correctness of such a characterization as we have in the representativeness of our sample and our ability to say when particular expressions are

meaningful (sentences). What clear and analogous task and test exist for semantics?[5]

We decided a while back not to assume that parts of sentences have meanings except in the ontologically neutral sense of making a systematic contribution to the meaning of the sentences in which they occur. Since postulating meanings has netted nothing, let us return to that insight. One direction in which it points is a certain holistic view of meaning. If sentences depend for their meaning on their structure, and we understand the meaning of each item in the structure only as an abstraction from the totality of sentences in which it features, then we can give the meaning of any sentence (or word) only by giving the meaning of every sentence (and word) in the language. Frege said that only in the context of a sentence does a word have meaning; in the same vein he might have added that only in the context of the language does a sentence (and therefore a word) have meaning.

This degree of holism was already implicit in the suggestion that an adequate theory of meaning must entail *all* sentences of the form '*s* means *m*'. But now, having found no more help in meanings of sentences than in meanings of words, let us ask whether we can get rid of the troublesome singular terms supposed to replace '*m*' and to refer to meanings. In a way, nothing could be easier: just write '*s* means that *p*', and imagine '*p*' replaced by a sentence. Sentences, as we have seen, cannot name meanings, and sentences with 'that' prefixed are not names at all, unless we decide so. It looks as though we are in trouble on another count, however, for it is reasonable to expect that in wrestling with the logic of the apparently non-extensional 'means that' we will encounter problems as hard as, or perhaps identical with, the problems our theory is out to solve.

The only way I know to deal with this difficulty is simple, and radical. Anxiety that we are enmeshed in the intensional springs from using the words 'means that' as filling between description of

[5] For a recent statement of the role of semantics in linguistics, see Noam Chomsky, 'Topics in the Theory of Generative Grammar'. In this article, Chomsky (1) emphasizes the central importance of semantics in linguistic theory, (2) argues for the superiority of transformational grammars over phrase-structure grammars largely on the grounds that, although phrase-structure grammars may be adequate to define sentencehood for (at least) some natural languages, they are inadequate as a foundation for semantics, and (3) comments repeatedly on the 'rather primitive state' of the concepts of semantics and remarks that the notion of semantic interpretation 'still resists any deep analysis'.

sentence and sentence, but it may be that the success of our venture depends not on the filling but on what it fills. The theory will have done its work if it provides, for every sentence *s* in the language under study, a matching sentence (to replace '*p*') that, in some way yet to be made clear, 'gives the meaning' of *s*. One obvious candidate for matching sentence is just *s* itself, if the object language is contained in the metalanguage; otherwise a translation of *s* in the metalanguage. As a final bold step, let us try treating the position occupied by '*p*' extensionally: to implement this, sweep away the obscure 'means that', provide the sentence that replaces '*p*' with a proper sentential connective, and supply the description that replaces '*s*' with its own predicate. The plausible result is

(*T*) *s* is *T* if and only if *p*.

What we require of a theory of meaning for a language *L* is that without appeal to any (further) semantical notions it place enough restrictions on the predicate 'is *T*' to entail all sentences got from schema *T* when '*s*' is replaced by a structural description of a sentence of *L* and '*p*' by that sentence.

Any two predicates satisfying this condition have the same extension,[6] so if the metalanguage is rich enough, nothing stands in the way of putting what I am calling a theory of meaning into the form of an explicit definition of a predicate 'is *T*'. But whether explicitly defined or recursively characterized, it is clear that the sentences to which the predicate 'is *T*' applies will be just the true sentences of *L*, for the condition we have placed on satisfactory theories of meaning is in essence Tarski's Convention *T* that tests the adequacy of a formal semantical definition of truth.[7]

The path to this point has been tortuous, but the conclusion may be stated simply: a theory of meaning for a language *L* shows 'how the meanings of sentences depend upon the meanings of words' if it contains a (recursive) definition of truth-in-L. And, so far at least, we have no other idea how to turn the trick. It is worth emphasizing that the concept of truth played no ostensible role in stating our original problem. That problem, upon refinement, led to the view that an adequate theory of meaning must characterize a predicate meeting certain conditions. It was in the nature of a discovery that

[6] Assuming, of course, that the extension of these predicates is limited to the sentences of *L*.

[7] A. Tarski, 'The Concept of Truth in Formalized Languages'.

such a predicate would apply exactly to the true sentences. I hope that what I am saying may be described in part as defending the philosophical importance of Tarski's semantical concept of truth. But my defence is only distantly related, if at all, to the question whether the concept Tarski has shown how to define is the (or a) philosophically interesting conception of truth, or the question whether Tarski has cast any light on the ordinary use of such words as 'true' and 'truth'. It is a misfortune that dust from futile and confused battles over these questions has prevented those with a theoretical interest in language—philosophers, logicians, psychologists, and linguists alike—from seeing in the semantical concept of truth (under whatever name) the sophisticated and powerful foundation of a competent theory of meaning.

There is no need to suppress, of course, the obvious connection between a definition of truth of the kind Tarski has shown how to construct, and the concept of meaning. It is this: the definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give truth conditions is a way of giving the meaning of a sentence. To know the semantic concept of truth for a language is to know what it is for a sentence—any sentence—to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language. This at any rate is my excuse for a feature of the present discussion that is apt to shock old hands; my freewheeling use of the word 'meaning', for what I call a theory of meaning has after all turned out to make no use of meanings, whether of sentences or of words. Indeed, since a Tarski-type truth definition supplies all we have asked so far of a theory of meaning, it is clear that such a theory falls comfortably within what Quine terms the 'theory of reference' as distinguished from what he terms the 'theory of meaning'. So much to the good for what I call a theory of meaning, and so much, perhaps, against my so calling it.[8]

A theory of meaning (in my mildly perverse sense) is an empirical theory, and its ambition is to account for the workings of a natural language. Like any theory, it may be tested by comparing some of its consequences with the facts. In the present case this is easy, for the

[8] But Quine may be quoted in support of my usage: '. . . in point of *meaning* . . . a word may be said to be determined to whatever extent the truth or falsehood of its contexts is determined.' ('Truth by Convention', 82.) Since a truth definition determines the truth value of every sentence in the object language (relative to a sentence in the metalanguage), it determines the meaning of every word and sentence. This would seem to justify the title Theory of Meaning.

theory has been characterized as issuing in an infinite flood of sentences each giving the truth conditions of a sentence; we only need to ask, in sample cases, whether what the theory avers to be the truth conditions for a sentence really are. A typical test case might involve deciding whether the sentence 'Snow is white' *is* true if and only if snow is white. Not all cases will be so simple (for reasons to be sketched), but it is evident that this sort of test does not invite counting noses. A sharp conception of what constitutes a theory in this domain furnishes an exciting context for raising deep questions about when a theory of language is correct and how it is to be tried. But the difficulties are theoretical, not practical. In application, the trouble is to get a theory that comes close to working; anyone can tell whether it is right.[9] One can see why this is so. The theory reveals nothing new about the conditions under which an individual sentence is true; it does not make those conditions any clearer than the sentence itself does. The work of the theory is in relating the known truth conditions of each sentence to those aspects ('words') of the sentence that recur in other sentences, and can be assigned identical roles in other sentences. Empirical power in such a theory depends on success in recovering the structure of a very complicated ability—the ability to speak and understand a language. We can tell easily enough when particular pronouncements of the theory comport with our understanding of the language; this is consistent with a feeble insight into the design of the machinery of our linguistic accomplishments.

The remarks of the last paragraph apply directly only to the special case where it is assumed that the language for which truth is being characterized is part of the language used and understood by the characterizer. Under these circumstances, the framer of a theory will as a matter of course avail himself when he can of the built-in convenience of a metalanguage with a sentence guaranteed equivalent to each sentence in the object language. Still, this fact ought not to con us into thinking a theory any more correct that entails '"Snow is white" is true if and only if snow is white' than one that entails instead:

(*S*) 'Snow is white' is true if and only if grass is green,

[9] To give a single example: it is clearly a count in favour of a theory that it entails '"Snow is white" is true if and only if snow is white'. But to contrive a theory that entails this (and works for all related sentences) is not trivial. I do not know a wholly satisfactory theory that succeeds with this very case (the problem of 'mass terms').

provided, of course, we are as sure of the truth of (*S*) as we are of that of its more celebrated predecessor. Yet (*S*) may not encourage the same confidence that a theory that entails it deserves to be called a theory of meaning.

The threatened failure of nerve may be counteracted as follows. The grotesqueness of (*S*) is in itself nothing against a theory of which it is a consequence, provided the theory gives the correct results for every sentence (on the basis of its structure, there being no other way). It is not easy to see how (*S*) could be party to such an enterprise, but if it were—if, that is, (*S*) followed from a characterization of the predicate 'is true' that led to the invariable pairing of truths with truths and falsehoods with falsehoods—then there would not, I think, be anything essential to the idea of meaning that remained to be captured.[10]

What appears to the right of the biconditional in sentences of the form '*s* is true if and only if *p*' when such sentences are consequences of a theory of truth plays its role in determining the meaning of *s* not by pretending synonymy but by adding one more brush-stroke to the picture which, taken as a whole, tells what there is to know of the meaning of *s*; this stroke is added by virtue of the fact that the sentence that replaces '*p*' is true if and only if *s* is.

It may help to reflect that (*S*) is acceptable, if it is, because we are independently sure of the truth of 'Snow is white' and 'Grass is green'; but in cases where we are unsure of the truth of a sentence, we can have confidence in a characterization of the truth predicate only if it pairs that sentence with one we have good reason to believe equivalent. It would be ill advised for someone who had any doubts about the colour of snow or grass to accept a theory that yielded (*S*), even if his doubts were of equal degree, unless he thought the colour of the one was tied to the colour of the other.[11] Omniscience can

[10] Critics have often failed to notice the essential proviso mentioned in this paragraph. The point is that (*S*) could not belong to any reasonably simple theory that also gave the right truth conditions for 'That is snow' and 'This is white'. (See the discussion of indexical expressions below.) [Footnote added in 1982.]

[11] This paragraph is confused. What it should say is that sentences of the theory are empirical generalizations about speakers, and so must not only be true but also lawlike. (*S*) presumably is not a law, since it does not support appropriate counterfactuals. It's also important that the evidence for accepting the (time and speaker relativized) truth conditions for 'That is snow' is based on the causal connection between a speaker's assent to the sentence and the demonstrative presentation of snow. For further discussion see Essay 12. [Footnote added in 1982.]

obviously afford more bizzare theories of meaning than ignorance; but then, omniscience has less need of communication.

It must be possible, of course, for the speaker of one language to construct a theory of meaning for the speaker of another, though in this case the empirical test of the correctness of the theory will no longer be trivial. As before, the aim of theory will be an infinite correlation of sentences alike in truth. But this time the theory-builder must not be assumed to have direct insight into likely equivalences between his own tongue and the alien. What he must do is find out, however he can, what sentences the alien holds true in his own tongue (or better, to what degree he holds them true). The linguist then will attempt to construct a characterization of truth-for-the-alien which yields, so far as possible, a mapping of sentences held true (or false) by the alien on to sentences held true (or false) by the linguist. Supposing no perfect fit is found, the residue of sentences held true translated by sentences held false (and vice versa) is the margin for error (foreign or domestic). Charity in interpreting the words and thoughts of others is unavoidable in another direction as well: just as we must maximize agreement, or risk not making sense of what the alien is talking about, so we must maximize the self-consistency we attribute to him, on pain of not understanding *him*. No single principle of optimum charity emerges; the constraints therefore determine no single theory. In a theory of radical translation (as Quine calls it) there is no completely disentangling questions of what the alien means from questions of what he believes. We do not know what someone means unless we know what he believes; we do not know what someone believes unless we know what he means. In radical interpretation we are able to break into this circle, if only incompletely, because we can sometimes tell that a person accedes to a sentence we do not understand.[12]

In the past few pages I have been asking how a theory of meaning that takes the form of a truth definition can be empirically tested, and have blithely ignored the prior question whether there is any serious chance such a theory can be given for a natural language. What are the prospects for a formal semantical theory of a natural

[12] This sketch of how a theory of meaning for an alien tongue can be tested obviously owes it inspiration to Quine's account of radical translation in Chapter II of *Word and Object*. In suggesting that an acceptable theory of radical translation take the form of a recursive characterization of truth, I go beyond Quine. Toward the end of this paper, in the discussion of demonstratives, another strong point of agreement will turn up.

language? Very poor, according to Tarski; and I believe most logicians, philosophers of language, and linguists agree.[13] Let me do what I can to dispel the pessimism. What I can in a general and programmatic way, of course, for here the proof of the pudding will certainly be in the proof of the right theorems.

Tarski concludes the first section of his classic essay on the concept of truth in formalized languages with the following remarks, which he italicizes:

> . . . *The very possibility of a consistent use of the expression 'true sentence' which is in harmony with the laws of logic and the spirit of everyday language seems to be very questionable, and consequently the same doubt attaches to the possibility of constructing a correct definition of this expression.* (165)

Late in the same essay, he returns to the subject:

> . . . the concept of truth (as well as other semantical concepts) when applied to colloquial language in conjunction with the normal laws of logic leads inevitably to confusions and contradictions. Whoever wishes, in spite of all difficulties, to pursue the semantics of colloquial language with the help of exact methods will be driven first to undertake the thankless task of a reform of this language. He will find it necessary to define its structure, to overcome the ambiguity of the terms which occur in it, and finally to split the language into a series of languages of greater and greater extent, each of which stands in the same relation to the next in which a formalized language stands to its metalanguage. It may, however be doubted whether the language of everyday life, after being 'rationalized' in this way, would still preserve its naturalness and whether it would not rather take on the characteristic features of the formalized languages. (267)

Two themes emerge: that the universal character of natural languages leads to contradiction (the semantic paradoxes), and that natural languages are too confused and amorphous to permit the direct application of formal methods. The first point deserves a serious answer, and I wish I had one. As it is, I will say only why I think we are justified in carrying on without having disinfected this particular source of conceptual anxiety. The semantic paradoxes arise when the range of the quantifiers in the object language is too generous in certain ways. But it is not really clear how unfair to Urdu or to Wendish it would be to view the range of their quantifiers

[13] So far as I am aware, there has been very little discussion of whether a formal truth definition can be given for a natural language. But in a more general vein, several people have urged that the concepts of formal semantics be applied to natural language. See, for example, the contributions of Yehoshua Bar-Hillel and Evert Beth to *The Philosophy of Rudolph Carnap*, and Bar-Hillel's 'Logical Syntax and Semantics'.

as insufficient to yield an explicit definition of 'true-in-Urdu' or 'true-in-Wendish'. Or, to put the matter in another, if not more serious way, there may in the nature of the case always be something we grasp in understanding the language of another (the concept of truth) that we cannot communicate to him. In any case, most of the problems of general philosophical interest arise within a fragment of the relevant natural language that may be conceived as containing very little set theory. Of course these comments do not meet the claim that natural languages are universal. But it seems to me that this claim, now that we know such universality leads to paradox, is suspect.

Tarski's second point is that we would have to reform a natural language out of all recognition before we could apply formal semantical methods. If this is true, it is fatal to my project, for the task of a theory of meaning as I conceive it is not to change, improve, or reform a language, but to describe and understand it. Let us look at the positive side. Tarski has shown the way to giving a theory for interpreted formal languages of various kinds; pick one as much like English as possible. Since this new language has been explained in English and contains much English we not only may, but I think must, view it as part of English for those who understand it. For this fragment of English we have, *ex hypothesi*, a theory of the required sort. Not only that, but in interpreting this adjunct of English in old English we necessarily gave hints connecting old and new. Wherever there are sentences of old English with the same truth conditions as sentences in the adjunct we may extend the theory to cover them. Much of what is called for is to mechanize as far as possible what we now do by art when we put ordinary English into one or another canonical notation. The point is not that canonical notation is better than the rough original idiom, but rather that if we know what idiom the canonical notation is canonical *for*, we have as good a theory for the idiom as for its kept companion.

Philosophers have long been at the hard work of applying theory to ordinary language by the device of matching sentences in the vernacular with sentences for which they have a theory. Frege's massive contribution was to show how 'all', 'some', 'every', 'each', 'none', and associated pronouns, in some of their uses, could be tamed; for the first time, it was possible to dream of a formal semantics for a significant part of a natural language. This dream came true in a sharp way with the work of Tarski. It would be a

shame to miss the fact that as a result of these two magnificent achievements, Frege's and Tarski's, we have gained a deep insight into the structure of our mother tongues. Philosophers of a logical bent have tended to start where the theory was and work out towards the complications of natural language. Contemporary linguists, with an aim that cannot easily be seen to be different, start with the ordinary and work toward a general theory. If either party is successful, there must be a meeting. Recent work by Chomsky and others is doing much to bring the complexities of natural languages within the scope of serious theory. To give an example: suppose success in giving the truth conditions for some significant range of sentences in the active voice. Then with a formal procedure for transforming each such sentence into a corresponding sentence in the passive voice, the theory of truth could be extended in an obvious way to this new set of sentences.[14]

One problem touched on in passing by Tarski does not, at least in all its manifestations, have to be solved to get ahead with theory: the existence in natural languages of 'ambiguous terms'. As long as ambiguity does not affect grammatical form, and can be translated, ambiguity for ambiguity, into the metalanguage, a truth definition will not tell us any lies. The chief trouble, for systematic semantics, with the phrase 'believes that' in English lies not in its vagueness, ambiguity, or unsuitability for incorporation in a serious science: let our metalanguage be English, and all *these* problems will be carried without loss or gain into the metalanguage. But the central problem of the logical grammar of 'believes that' will remain to haunt us.

The example is suited to illustrating another, and related, point, for the discussion of belief sentences has been plagued by failure to

[14] The *rapprochement* I prospectively imagine between transformational grammar and a sound theory of meaning has been much advanced by a recent change in the conception of transformational grammar described by Chomsky in the article referred to above (note 5). The structures generated by the phrase-structure part of the grammar, it has been realized for some time, are those suited to semantic interpretation; but this view is inconsistent with the idea, held by Chomsky until recently, that recursive operations are introduced only by the transformation rules. Chomsky now believes the phrase-structure rules are recursive. Since languages to which formal semantic methods directly and naturally apply are ones for which a (recursive) phrase-structure grammar is appropriate, it is clear that Chomsky's present picture of the relation between the structures generated by the phrase-structure part of the grammar, and the sentences of the language, is very much like the picture many logicians and philosophers have had of the relation between the richer formalized languages and ordinary language. (In these remarks I am indebted to Bruce Vermazen.)

observe a fundamental distinction between tasks: uncovering the logical grammar or form of sentences (which is in the province of a theory of meaning as I construe it), and the analysis of individual words or expressions (which are treated as primitive by the theory). Thus Carnap, in the first edition of *Meaning and Necessity*, suggested we render 'John believes that the earth is round' as 'John responds affirmatively to "the earth is round" as an English sentence'. He gave this up when Mates pointed out that John might respond affirmatively to one sentence and not to another no matter how close in meaning.[15] But there is a confusion here from the start. The semantic structure of a belief sentence, according to this idea of Carnap's, is given by a three-place predicate with places reserved for expressions referring to a person, a sentence, and a language. It is a different sort of problem entirely to attempt an analysis of this predicate, perhaps along behaviouristic lines. Not least among the merits of Tarski's conception of a theory of truth is that the purity of method it demands of us follows from the formulation of the problem itself, not from the self-imposed restraint of some adventitious philosophical puritanism.

I think it is hard to exaggerate the advantages to philosophy of language of bearing in mind this distinction between questions of logical form or grammar, and the analysis of individual concepts. Another example may help advertise the point.

If we suppose questions of logical grammar settled, sentences like 'Bardot is good' raise no special problems for a truth definition. The deep differences between descriptive and evaluative (emotive, expressive, etc.) terms do not show here. Even if we hold there is some important sense in which moral or evaluative sentences do not have a truth value (for example, because they cannot be verified), we ought not to boggle at '"Bardot is good" is true if and only if Bardot is good'; in a theory of truth, this consequence should follow with the rest, keeping track, as must be done, of the semantic location of such sentences in the language as a whole—of their relation to generalizations, their role in such compound sentences as 'Bardot is good and Bardot is foolish', and so on. What is special to evaluative words is simply not touched: the mystery is transferred from the word 'good' in the object language to its translation in the metalanguage.

[15] B. Mates, 'Synonymity

But 'good' as it features in 'Bardot is a good actress' is another matter. The problem is not that the translation of this sentence is not in the metalanguage—let us suppose it is. The problem is to frame a truth definition such that '"Bardot is a good actress" is true if and only if Bardot is a good actress'—and all other sentences like it—are consequences. Obviously 'good actress' does not mean 'good and an actress'. We might think of taking 'is a good actress' as an unanalysed predicate. This would obliterate all connection between 'is a good actress' and 'is a good mother', and it would give us no excuse to think of 'good', in these uses, as a word or semantic element. But worse, it would bar us from framing a truth definition at all, for there is no end to the predicates we would have to treat as logically simple (and hence accommodate in separate clauses in the definition of satisfaction): 'is a good companion to dogs', 'is a good 28-years old conversationalist', and so forth. The problem is not peculiar to the case: it is the problem of attributive adjectives generally.

It is consistent with the attitude taken here to deem it usually a strategic error to undertake philosophical analysis of words or expressions which is not preceded by or at any rate accompanied by the attempt to get the logical grammar straight. For how can we have any confidence in our analyses of words like 'right', 'ought', 'can', and 'obliged', or the phrases we use to talk of actions, events, and causes, when we do not know what (logical, semantical) parts of speech we have to deal with? I would say much the same about studies of the 'logic' of these and other words, and the sentences containing them. Whether the effort and ingenuity that have gone into the study of deontic logics, modal logics, imperative and erotetic logics have been largely futile or not cannot be known until we have acceptable semantic analyses of the sentences such systems purport to treat. Philosophers and logicians sometimes talk or work as if they were free to choose between, say, the truth-functional conditional and others, or free to introduce non-truth-functional sentential operators like 'Let it be the case that' or 'It ought to be the case that'. But in fact the decision is crucial. When we depart from idioms we can accommodate in a truth definition, we lapse into (or create) language for which we have no coherent semantical account—that is, no account at all of how such talk can be integrated into the language as a whole.

To return to our main theme: we have recognized that a theory of

the kind proposed leaves the whole matter of what individual words mean exactly where it was. Even when the metalanguage is different from the object language, the theory exerts no pressure for improvement, clarification, or analysis of individual words, except when, by accident of vocabulary, straightforward translation fails. Just as synonomy, as between expressions, goes generally untreated, so also synonomy of sentences, and analyticity. Even such sentences as 'A vixen is a female fox' bear no special tag unless it is our pleasure to provide it. A truth definition does not distinguish between analytic sentences and others, except for sentences that owe their truth to the presence alone of the constants that give the theory its grip on structure: the theory entails not only that these sentences are true but that they will remain true under all significant rewritings of their non-logical parts. A notion of logical truth thus given limited application, related notions of logical equivalence and entailment will tag along. It is hard to imagine how a theory of meaning could fail to read a logic into its object language to this degree; and to the extent that it does, our intuitions of logical truth, equivalence, and entailment may be called upon in constructing and testing the theory.

I turn now to one more, and very large, fly in the ointment: the fact that the same sentence may at one time or in one mouth be true and at another time or in another mouth be false. Both logicians and those critical of formal methods here seem largely (though by no means universally) agreed that formal semantics and logic are incompetent to deal with the disturbances caused by demonstratives. Logicians have often reacted by downgrading natural language and trying to show how to get along without demonstratives; their critics react by downgrading logic and formal semantics. None of this can make me happy: clearly demonstratives cannot be eliminated from a natural language without loss or radical change, so there is no choice but to accommodate theory to them.

No logical errors result if we simply treat demonstratives as constants;[16] neither do any problems arise for giving a semantic truth definition. '"I am wise" is true if and only if I am wise', with its bland ignoring of the demonstrative element in 'I' comes off the assembly line along with '"Socrates is wise" is true if and only if Socrates is wise' with *its* bland indifference to the demonstrative element in 'is wise' (the tense).

[16] See W. V. Quine, *Methods of Logic*, 8.

What suffers in this treatment of demonstratives is not the definition of a truth predicate, but the plausibility of the claim that what has been defined is truth. For this claim is acceptable only if the speaker and circumstances of utterance of each sentence mentioned in the definition is matched by the speaker and circumstances of utterance of the truth definition itself. It could also be fairly pointed out that part of understanding demonstratives is knowing the rules by which they adjust their reference to circumstance; assimilating demonstratives to constant terms obliterates this feature. These complaints can be met, I think, though only by a fairly far-reaching revision in the theory of truth. I shall barely suggest how this could be done, but bare suggestion is all that is needed: the idea is technically trivial, and in line with work being done on the logic of the tenses.[17]

We could take truth to be a property, not of sentences, but of utterances, or speech acts, or ordered triples of sentences, times, and persons; but it is simplest just to view truth as a relation between a sentence, a person, and a time. Under such treatment, ordinary logic as now read applies as usual, but only to sets of sentences relativized to the same speaker and time; further logical relations between sentences spoken at different times and by different speakers may be articulated by new axioms. Such is not my concern. The theory of meaning undergoes a systematic but not puzzling change; corresponding to each expression with a demonstrative element there must in the theory be a phrase that relates the truth conditions of sentences in which the expression occurs to changing times and speakers. Thus the theory will entail sentences like the following:

'I am tired' is true as (potentially) spoken by p at t if and only if p is tired at t.

'That book was stolen' is true as (potentially) spoken by p at t if and only if the book demonstrated by p at t is stolen prior to t.[18]

Plainly, this course does not show how to eliminate demonstratives; for example, there is no suggestion that 'the book demonstrated by the speaker' can be substituted ubiquitously for 'that book' *salva veritate*. The fact that demonstratives are amenable to

[17] This claim has turned out to be naïvely optimistic. For some serious work on the subject, see S. Weinstein, 'Truth and Demonstratives'. [Note added in 1982.]

[18] There is more than an intimation of this approach to demonstratives and truth in J. L. Austin, 'Truth'.

formal treatment ought greatly to improve hopes for a serious semantics of natural language, for it is likely that many outstanding puzzles, such as the analysis of quotations or sentences about propositional attitudes, can be solved if we recognize a concealed demonstrative construction.

Now that we have relativized truth to times and speakers, it is appropriate to glance back at the problem of empirically testing a theory of meaning for an alien tongue. The essence of the method was, it will be remembered, to correlate held-true sentences with held-true sentences by way of a truth definition, and within the bounds of intelligible error. Now the picture must be elaborated to allow for the fact that sentences are true, and held true, only relative to a speaker and a time. Sentences with demonstratives obviously yield a very sensitive test of the correctness of a theory of meaning, and constitute the most direct link between language and the recurrent macroscopic objects of human interest and attention.[19]

In this paper I have assumed that the speakers of a language can effectively determine the meaning or meanings of an arbitrary expression (if it has a meaning), and that it is the central task of a theory of meaning to show how this is possible. I have argued that a characterization of a truth predicate describes the required kind of structure, and provides a clear and testable criterion of an adequate semantics for a natural language. No doubt there are other reasonable demands that may be put on a theory of meaning. But a theory that does no more than define truth for a language comes far closer to constituting a complete theory of meaning than superficial analysis might suggest; so, at least, I have urged.

Since I think there is no alternative, I have taken an optimistic and programmatic view of the possibilities for a formal characterization of a truth predicate for a natural language. But it must be allowed that a staggering list of difficulties and conundrums remains. To name a few: we do not know the logical form of counterfactual or subjunctive sentences; nor of sentences about probabilities and about causal relations; we have no good idea what the logical role of adverbs is, nor the role of attributive adjectives; we have no theory for mass terms like 'fire', 'water', and 'snow', nor for sentences about

[19] These remarks derive from Quine's idea that 'occasion sentences' (those with a demonstrative element) must play a central role in constructing a translation manual.

belief, perception, and intention, nor for verbs of action that imply purpose. And finally, there are all the sentences that seem not to have truth values at all: the imperatives, optatives, interrogatives, and a host more. A comprehensive theory of meaning for a natural language must cope successfully with each of these problems.[20]

[20] For attempted solutions to some of these problems see Essays 6–10 of *Essays on Actions and Events*, and Essays 6–8 of this book. There is further discussion in Essays 3, 4, 9, and 10, and reference to some progress in section 1 of Essay 9.

§ 2. THE OBJECTIVE PULL; OR, E PLURIBUS UNUM

'Ouch' is a one-word sentence which a man may volunteer from time to time by way of laconic comment on the passing show. The correct occasions of its use are those attended by painful stimulation. Such use of the word, like the correct use of language generally, is inculcated in the individual by training on the part of society; and society achieves this despite not sharing the individual's pain. Society's method is in principle that of rewarding the utterance of 'Ouch' when the speaker shows some further evidence of sudden discomfort, say a wince, or is actually seen to suffer violence, and of penalizing the utterance of 'Ouch' when the speaker is visibly untouched and his countenance unruffled.

For the man who has learned his language lesson, some of the stimuli evocative of 'Ouch' may be publicly visible blows and slashes, while others are hidden from the public eye in the depths of his bowels. Society, acting solely on overt manifestations, has been able to train the individual to say the socially proper thing in response even to socially undetectable stimulations. The trick has

QUINE, W.V.O.
WORD AND OBJECT

depended on prior concomitances between covert stimulation and overt behavior, notably the wincing instinct.

We can imagine a primitive use of 'Red' as a one-word sentence somewhat on a par with 'Ouch'. Just as 'Ouch' is the appropriate remark on the occasion of painful stimulation, so 'Red', under the usage which I am now imagining, is the appropriate remark on the occasion of those distinctive photochemical effects which are wrought in one's retina by the impact of red light. This time society's method of training consists in rewarding the utterance of 'Red' when the individual is seen looking at something red, and penalizing it when he is seen looking at something else.

Actually the uses of 'Red' are less simple. Commonly 'red', unlike 'ouch', turns up as a fragment of longer sentences. Moreover, even when 'Red' is used by itself as a one-word sentence, what evokes it is usually not the mere apprehension of something red; more commonly there has been a verbal stimulus, in the form of a question. But let us keep for a moment to the fictitious usage described in the preceding paragraph; for it, by its similarity to 'Ouch', will help to bring out also a certain contrast.

The critic, society's agent, approves the subject's utterance of 'Red' by observing the subject and his viewed object and finding the latter red. In part, therefore, the critic's cue is red irradiation of his own retina. A partial symmetry obtains between the subject's cue for utterance and the critic's cue for approval in the case of 'Red', which, happily for the critic, was lacking in the case of 'Ouch'. The partial symmetry in the one case, and the lack of it in the other, suggest a certain superficial sense in which 'Ouch' may be spoken of as more subjective in reference than 'Red'; 'Red' more objective than 'Ouch'.

Exceptions are possible on either side. If the critic and the subject are fighting a fire and are scorched by the same sudden gust, then the critic's approval of the subject's 'Ouch' does not differ significantly from the imagined case of 'Red'. Conversely, a critic may approve a subject's 'Red' on indirect evidence, failing to glimpse the object himself. If we call 'Ouch' more subjective than 'Red', we must be taken as alluding thereby only to the most characteristic learning situations. In the case of 'Red', typically one's mentor or critic sees red; in the case of 'Ouch', typically he does not get hurt.

'Ouch' is not independent of social training. One has only to

prick a foreigner to appreciate that it is an English word. But in its subjectivity it is a little unusual. Words being social tools, objectivity counts toward their survival. When a word has considerable currency despite the subjective twist, it may be expected, like the pronouns 'I' and 'you', to have a valuable social function of some exceptional sort. The survival value of 'Ouch', from a social point of view, is as a distress signal. And the word is of only marginal linguistic status, after all, being incapable of integration into longer sentences.

The usual premium on objectivity is well illustrated by 'square'. Each of a party of observers glances at a tile from his own vantage point and calls it square; and each of them has, as his retinal projection of the tile, a scalene quadrilateral which is geometrically dissimilar to everyone else's. The learner of 'square' has to take his chances with the rest of society, and he ends up using the word to suit. Association of 'square' with just the situations in which the retinal projection is square would be simpler to learn, but the more objective usage is, by its very intersubjectivity, what we tend to be exposed to and encouraged in.

In general, if a term is to be learned by induction from observed instances where it is applied, the instances have to resemble one another in two ways: they have to be enough alike from the learner's point of view, from occasion to occasion, to afford him a basis of similarity to generalize upon, and they have to be enough alike from simultaneous distinct points of view to enable the teacher and learner to share the appropriate occasions. A term restricted to squares normal to the line of sight would meet the first requirement only; a term applying to physical squares in all their scalene projections meets both. And it meets both in the same way, in that the points of view available to the learner from occasion to occasion are likewise the points of view available to teacher and learner on simultaneous occasions. Such is the way with terms for observable physical objects generally; and thus it is that such objects are focal to reference and thought.

'Red', unlike 'square', is a happy case where a nearly uniform stimulatory condition is shared by simultaneous observers. All the assembled retinas are irradiated by substantially the same red light, whereas no two of them receive geometrically similar projections of the square. The pull toward objectivity is thus a strong pull away from the subjectively simplest rule of association in the case of

'square', and much less so in the case of 'red'. Hence our readiness to think of color as more subjective than physical shape. But some pull of the same kind occurs even in the case of 'red', insofar as reflections from the environment cause the red object to cast somewhat different tints to different points of view. The objective pull will regiment all the responses still as 'red', by activating myriad corrective cues. These corrective cues are used unconsciously, such is the perfection of our socialization; a painter has even to school himself to set them aside when he tries to reproduce his true retinal intake.

The uniformity that unites us in communication and belief is a uniformity of resultant patterns overlying a chaotic subjective diversity of connections between words and experience. Uniformity comes where it matters socially; hence rather in point of intersubjectively conspicuous circumstances of utterance than in point of privately conspicuous ones. For an extreme illustration of the point, consider two men one of whom has normal color vision and the other of whom is color-blind as between red and green. Society has trained both men by the method noted earlier: rewarding the utterance of 'red' when the speaker is seen fixating something red, and penalizing it in the contrary case. Moreover the gross socially observable results are about alike: both men are pretty good about attributing 'red' to just the red things. But the private mechanisms by which the two men achieve these similar results are very different. The one man has learned 'red' in association with the regulation photochemical effect. The other man has painfully learned to be stimulated to 'red' by light in various wavelengths (red and green) in company with elaborate special combinations of supplementary conditions of intensity, saturation, shape, and setting, calculated e.g. to admit fire and sunsets and to exclude grass; to admit blossoms and exclude leaves; and to admit lobsters only after boiling.

Different persons growing up in the same language are like different bushes trimmed and trained to take the shape of identical elephants. The anatomical details of twigs and branches will fulfill the elephantine form differently from bush to bush, but the overall outward results are alike.

DAVIDSON, DONALD

IN:
INQUIRIES INTO TRUTH
AND MEANING

9 *Radical Interpretation*

Kurt utters the words 'Es regnet' and under the right conditions we know that he has said that it is raining. Having identified his utterance as intentional and linguistic, we are able to go on to interpret his words: we can say what his words, on that occasion, meant. What could we know that would enable us to do this? How could we come to know it? The first of these questions is not the same as the question what we *do* know that enables us to interpret the words of others. For there may easily be something we could know and don't, knowledge of which would suffice for interpretation, while on the other hand it is not altogether obvious that there is anything we actually know which plays an essential role in interpretation. The second question, how we could come to have knowledge that would serve to yield interpretations, does not, of course, concern the actual history of language acquisition. It is thus a doubly hypothetical question: given a theory that would make interpretation possible, what evidence plausibly available to a potential interpreter would support the theory to a reasonable degree? In what follows I shall try to sharpen these questions and suggest answers.

The problem of interpretation is domestic as well as foreign: it surfaces for speakers of the same language in the form of the question, how can it be determined that the language is the same? Speakers of the same language can go on the assumption that for them the same expressions are to be interpreted in the same way, but this does not indicate what justifies the assumption. All understanding of the speech of another involves radical interpretation. But it will help keep assumptions from going unnoticed to focus on cases

where interpretation is most clearly called for: interpretation in one idiom of talk in another.[1]

What knowledge would serve for interpretation? A short answer would be, knowledge of what each meaningful expression means. In German, those words Kurt spoke mean that it is raining and Kurt was speaking German. So in uttering the words 'Es regnet', Kurt said that it was raining. This reply does not, as might first be thought, merely restate the problem. For it suggests that in passing from a description that does not interpret (his uttering of the words 'Es regnet') to interpreting description (his saying that it is raining) we must introduce a machinery of words and expressions (which may or may not be exemplified in actual utterances), and this suggestion is important. But the reply is no further help, for it does not say what it is to know what an expression means.

There is indeed also the hint that corresponding to each meaningful expression that is an entity, its meaning. This idea, even if not wrong, has proven to be very little help: at best it hypostasizes the problem.

Disenchantment with meanings as implementing a viable account of communication or interpretation helps explain why some philosophers have tried to get along without, not only meanings, but any serious theory at all. It is tempting, when the concepts we summon up to try to explain interpretation turn out to be more baffling than the explanandum, to reflect that after all verbal communication consists in nothing more than elaborate disturbances in the air which form a causal link between the non-linguistic activities of human agents. But although interpretable speeches are nothing but (that is, identical with) actions performed with assorted non-linguistic intentions (to warn, control, amuse, distract, insult), and these actions are in turn nothing but (identical with) intentional movements of the lips and larynx, this observation takes us no distance towards an intelligible general account of what we might know that would allow us to redescribe uninterpreted utterances as the right interpreted ones.

Appeal to meanings leaves us stranded further than we started from the non-linguistic goings-on that must supply the evidential base for interpretation; the 'nothing but' attitude provides no clue as to how the evidence is related to what it surely is evident for.

[1] The term 'radical interpretation' is meant to suggest strong kinship with Quine's 'radical translation'. Kinship is not identity, however, and 'interpretation' in place of 'translation' marks one of the differences: a greater emphasis on the explicitly semantical in the former.

Other proposals for bridging the gap fall short in various ways. The 'causal' theories of Ogden and Richards and of Charles Morris attempted to analyse the meaning of sentences, taken one at a time, on the basis of behaviouristic data. Even if these theories had worked for the simplest sentences (which they clearly did not), they did not touch the problem of extending the method to sentences of greater complexity and abstractness. Theories of another kind start by trying to connect words rather than sentences with non-linguistic facts. This is promising because words are finite in number while sentences are not, and yet each sentence is no more than a concatenation of words: this offers the chance of a theory that interprets each of an infinity of sentences using only finite resources. But such theories fail to reach the evidence, for it seems clear that the semantic features of words cannot be explained directly on the basis of non-linguistic phenomena. The reason is simple. The phenomena to which we must turn are the extra-linguistic interests and activities that language serves, and these are served by words only in so far as the words are incorporated in (or on occasion happen to be) sentences. But then there is no chance of giving a foundational account of words before giving one of sentences.

For quite different reasons, radical interpretation cannot hope to take as evidence for the meaning of a sentence an account of the complex and delicately discriminated intentions with which the sentence is typically uttered. It is not easy to see how such an approach can deal with the structural, recursive feature of language that is essential to explaining how new sentences can be understood. But the central difficulty is that we cannot hope to attach a sense to the attribution of finely discriminated intentions independently of interpreting speech. The reason is not that we cannot ask necessary questions, but that interpreting an agent's intentions, his beliefs and his words are parts of a single project, no part of which can be assumed to be complete before the rest is. If this is right, we cannot make the full panoply of intentions and beliefs the evidential base for a theory of radical interpretation.

We are now in a position to say something more about what would serve to make interpretation possible. The interpreter must be able to understand any of the infinity of sentences the speaker might utter. If we are to state explicitly what the interpreter might know

that would enable him to do this, we must put it in finite form.[2] If this requirement is to be met, any hope of a universal method of interpretation must be abandoned. The most that can be expected is to explain how an interpreter could interpret the utterances of speakers of a single language (or a finite number of languages): it makes no sense to ask for a theory that would yield an explicit interpretation for any utterance in any (possible) language.

It is still not clear, of course, what it is for a theory to yield an explicit interpretation of an utterance. The formulation of the problem seems to invite us to think of the theory as the specification of a function taking utterances as arguments and having interpretations as values. But then interpretations would be no better than meanings and just as surely entities of some mysterious kind. So it seems wise to describe what is wanted of the theory without apparent reference to meanings or interpretations: someone who knows the theory can interpret the utterances to which the theory applies.

The second general requirement on a theory of interpretation is that it can be supported or verified by evidence plausibly available to an interpreter. Since the theory is general—it must apply to a potential infinity of utterances—it would be natural to think of evidence in its behalf as instances of particular interpretations recognized as correct. And this case does, of course, arise for the interpreter dealing with a language he already knows. The speaker of a language normally cannot produce an explicit finite theory for his own language, but he can test a proposed theory since he can tell whether it yields correct interpretations when applied to particular utterances.

In radical interpretation, however, the theory is supposed to supply an understanding of particular utterances that is not given in advance, so the ultimate evidence for the theory cannot be correct sample interpretations. To deal with the general case, the evidence must be of a sort that would be available to someone who does not already know how to interpret utterances the theory is designed to cover: it must be evidence that can be stated without essential use of such linguistic concepts as meaning, interpretation, synonymy, and the like.

Before saying what kind of theory I think will do the trick, I want

[2] See Essay 1.

to discuss a last alternative suggestion, namely that a method of translation, from the language to be interpreted into the language of the interpreter, is all the theory that is needed. Such a theory would consist in the statement of an effective method for going from an arbitrary sentence of the alien tongue to a sentence of a familiar language; thus it would satisfy the demand for a finitely stated method applicable to any sentence. But I do not think a translation manual is the best form for a theory of interpretation to take.[3]

When interpretation is our aim, a method of translation deals with a wrong topic, a relation between two languages, where what is wanted is an interpretation of one (in another, of course, but that goes without saying since any theory is in some language). We cannot without confusion count the language used in stating the theory as part of the subject matter of the theory unless we explicitly make it so. In the general case, a theory of translation involves three languages: the object language, the subject language, and the metalanguage (the languages from and into which translation proceeds, and the language of the theory, which says what expressions of the subject language translate which expressions of the object language). And in this general case, we can know which sentences of the subject language translate which sentences of the object language without knowing what any of the sentences of either language mean (in any sense, anyway, that would let someone who understood the theory interpret sentences of the object language). If the subject language happens to be identical with the language of the theory, then someone who understands the theory can no doubt use the translation manual to interpret alien utterances; but this is because he brings to bear two things he knows and that the theory does not state: the fact that the subject language is his own, and his knowledge of how to interpet utterances in his own language.

It is awkward to try to make explicit the assumption that a mentioned sentence belongs to one's own language. We could try, for example, '"Es regnet" in Kurt's language is translated as "It is raining" in mine', but the indexical self-reference is out of place in a theory that ought to work for any interpreter. If we decide to accept

[3] The idea of a translation manual with appropriate empirical constraints as a device for studying problems in the philosophy of language is, of course, Quine's. This idea inspired much of my thinking on the present subject, and my proposal is in important respects very close to Quine's. Since Quine did not intend to answer the questions I have set, the claim that the method of translation is not adequate as a solution to the problem of radical interpretation is not a criticism of any doctrine of Quine's.

this difficulty, there remains the fact that the method of translation leaves tacit and beyond the reach of theory what we need to know that allows us to interpret our own language. A theory of translation must read some sort of structure into sentences, but there is no reason to expect that it will provide any insight into how the meanings of sentences depend on their structure.

A satisfactory theory for interpreting the utterances of a language, our own included, will reveal significant semantic structure: the interpretation of utterances of complex sentences will systematically depend on the interpretation of utterances of simpler sentences, for example. Suppose we were to add to a theory of translation a satisfactory theory of interpretation for our own language. Then we would have exactly what we want, but in an unnecessarily bulky form. The translation manual churns out, for each sentence of the language to be translated, a sentence of the translator's language; the theory of interpretation then gives the interpretation of these familiar sentences. Clearly the reference to the home language is superfluous; it is an unneeded intermediary between interpretation and alien idiom. The only expressions a theory of interpretation has to mention are those belonging to the language to be interpreted.

A theory of interpretation for an object language may then be viewed as the result of the merger of a structurally revealing theory of interpretation for a known language, and a system of translation from the unknown language into the known. The merger makes all reference to the known language otiose; when this reference is dropped, what is left is a structurally revealing theory of interpretation for the object language—couched, of course, in familiar words. We have such theories, I suggest, in theories of truth of the kind Tarski first showed how to give.[4]

What characterizes a theory of truth in Tarski's style is that it entails, for every sentence *s* of the object language, a sentence of the form:

s is true (in the object language) if and only if *p*.

Instances of the form (which we shall call T-sentences) are obtained by replacing '*s*' by a canonical description of *s*, and '*p*' by a translation of *s*. The important undefined semantical notion in the theory is that of *satisfaction* which relates sentences, open or closed,

[4] A. Tarski, 'The Concept of Truth in Formalized Languages'.

to infinite sequences of objects, which may be taken to belong to the range of the variables of the object language. The axioms, which are finite in number, are of two kinds: some give the conditions under which a sequence satisfies a complex sentence on the basis of the conditions of satisfaction of simpler sentences, others give the conditions under which the simplest (open) sentences are satisfied. Truth is defined for closed sentences in terms of the notion of satisfaction. A recursive theory like this can be turned into an explicit definition along familiar lines, as Tarski shows, provided the language of the theory contains enough set theory; but we shall not be concerned with this extra step.

Further complexities enter if proper names and functional expressions are irreducible features of the object language. A trickier matter concerns indexical devices. Tarski was interested in formalized languages containing no indexical or demonstrative aspects. He could therefore treat sentences as vehicles of truth; the extension of the theory to utterances is in this case trivial. But natural languages are indispensably replete with indexical features, like tense, and so their sentences may vary in truth according to time and speaker. The remedy is to characterize truth for a language relative to a time and a speaker. The extension to utterances is again straightforward.[5]

What follows is a defence of the claim that a theory of truth, modified to apply to a natural language, can be used as a theory of interpretation. The defence will consist in attempts to answer three questions:

1. Is it reasonable to think that a theory of truth of the sort described can be given for a natural language?
2. Would it be possible to tell that such a theory was correct on the basis of evidence plausibly available to an interpreter with no prior knowledge of the language to be interpreted?
3. If the theory were known to be true, would it be possible to interpret utterances of speakers of the language?

The first question is addressed to the assumption that a theory of truth can be given for a natural language: the second and third questions ask whether such a theory would satisfy the further demands we have made on a theory of interpretation.

[5] For a discussion of how a theory of truth can handle demonstratives and how Convention T must be modified, see S. Weinstein, 'Truth and Demonstratives'.

1. *Can a theory of truth be given for a natural language?*

It will help us to appreciate the problem to consider briefly the case where a significant fragment of a language (plus one or two semantical predicates) is used to state its own theory of truth. According to Tarski's Convention T, it is a test of the adequacy of a theory that it entails all the T-sentences. This test apparently cannot be met without assigning something very much like a standard quantificational form to the sentences of the language, and appealing, in the theory, to a relational notion of satisfaction.[6] But the striking thing about T-sentences is that whatever machinery must operate to produce them, and whatever ontological wheels must turn, in the end a T-sentence states the truth conditions of a sentence using resources no richer than, because the same as, those of the sentence itself. Unless the original sentence mentions possible worlds, intensional entities, properties, or propositions, the statement of its truth conditions does not.

There is no equally simple way to make the analogous point about an alien language without appealing, as Tarski does, to an unanalysed notion of translation. But what we can do for our own language we ought to be able to do for another; the problem, it will turn out, will be to know that we are doing it.

The restriction imposed by demanding a theory that satisfies Convention T seems to be considerable: there is no generally accepted method now known for dealing, within the restriction, with a host of problems, for example, sentences that attribute attitudes, modalities, general causal statements, counterfactuals, attributive adjectives, quantifiers like 'most', and so on. On the other hand, there is what seems to me to be fairly impressive progress. To mention some examples, there is the work of Tyler Burge on proper names,[7] Gilbert Harman on 'ought',[8] John Wallace on mass terms and comparatives,[9] and there is my own work on attributions of attitudes and performatives,[10] on adverbs, events, and singular causal statements,[11] and on quotation.[12]

If we are inclined to be pessimistic about what remains to be done

[6] See J. Wallace, 'On the Frame of Reference', and Essay 3.
[7] T. Burge, 'Reference and Proper Names'.
[8] G. Harman, 'Moral Relativism Defended'.
[9] J. Wallace, 'Positive, Comparative, Superlative'.
[10] See Essays 7 and 8.
[11] See Essays 6–10 in *Essays on Actions and Events*.
[12] See Essay 6.

(or some of what has been done!), we should think of Frege's magnificent accomplishment in bringing what Dummett calls 'multiple generality' under control.[13] Frege did not have a theory of truth in Tarski's sense in mind, but it is obvious that he sought, and found, structures of a kind for which a theory of truth can be given.

The work of applying a theory of truth in detail to a natural language will in practice almost certainly divide into two stages. In the first stage, truth will be characterized, not for the whole language, but for a carefully gerrymandered part of the language. This part, though no doubt clumsy grammatically, will contain an infinity of sentences which exhaust the expressive power of the whole language. The second part will match each of the remaining sentences to one or (in the case of ambiguity) more than one of the sentences for which truth has been characterized. We may think of the sentences to which the first stage of the theory applies as giving the logical form, or deep structure, of all sentences.

2. *Can a theory of truth be verified by appeal to evidence available before interpretation has begun?*

Convention T says that a theory of truth is satisfactory if it generates a T-sentence for each sentence of the object language. It is enough to demonstrate that a theory of truth is empirically correct, then, to verify that the T-sentences are true (in practice, an adequate sample will confirm the theory to a reasonable degree). T-sentences mention only the closed sentences of the language, so the relevant evidence can consist entirely of facts about the behaviour and attitudes of speakers in relation to sentences (no doubt by way of utterances). A workable theory must, of course, treat sentences as concatenations of expressions of less than sentential length, it must introduce semantical notions like satisfaction and reference, and it must appeal to an ontology of sequences and the objects ordered by the sequences. All this apparatus is properly viewed as theoretical construction, beyond the reach of direct verification. It has done its work provided only it entails testable results in the form of T-sentences, and these make no mention of the machinery. A theory of truth thus reconciles the demand for a theory that articulates

[13] M. Dummett, *Frege: Philosophy of Language*.

grammatical structure with the demand for a theory that can be tested only by what it says about sentences.

In Tarski's work, T-sentences are taken to be true because the right branch of the biconditional is assumed to be a translation of the sentence truth conditions for which are being given. But we cannot assume in advance that correct translation can be recognized without pre-empting the point of radical interpretation; in empirical applications, we must abandon the assumption. What I propose is to reverse the direction of explanation: assuming translation, Tarski was able to define truth; the present idea is to take truth as basic and to extract an account of translation or interpretation. The advantages, from the point of view of radical interpretation, are obvious. Truth is a single property which attaches, or fails to attach, to utterances, while each utterance has its own interpretation; and truth is more apt to connect with fairly simple attitudes of speakers.

There is no difficulty in rephrasing Convention T without appeal to the concept of translation: an acceptable theory of truth must entail, for every sentence *s* of the object language, a sentence of the form: *s* is true if and only if *p*, where '*p*' is replaced by any sentence that is true if and only if *s* is. Given this formulation, the theory is tested by evidence that T-sentences are simply true; we have given up the idea that we must also tell whether what replaces '*p*' translates *s*. It might seem that there is no chance that if we demand so little of T-sentences, a theory of interpretation will emerge. And of course this would be so if we took the T-sentences in isolation. But the hope is that by putting appropriate formal and empirical restrictions on the theory as a whole, individual T-sentences will in fact serve to yield interpretations.[14]

We have still to say what evidence is available to an interpreter—evidence, we now see, that T-sentences are true. The evidence cannot consist in detailed descriptions of the speaker's beliefs and intentions, since attributions of attitudes, at least where subtlety is required, demand a theory that must rest on much the same evidence as interpretation. The interdependence of belief and meaning is evident in this way: a speaker holds a sentence to be true because of what the sentence (in his language) means, and because of what he believes. Knowing that he holds the sentence to be true, and knowing the meaning, we can infer his belief; given enough

[14] For essential qualifications, see footnote 11 of Essay 2.

information about his beliefs, we could perhaps infer the meaning. But radical interpretation should rest on evidence that does not assume knowledge of meanings or detailed knowledge of beliefs.

A good place to begin is with the attitude of holding a sentence true, of accepting it as true. This is, of course, a belief, but it is a single attitude applicable to all sentences, and so does not ask us to be able to make finely discriminated distinctions among beliefs. It is an attitude an interpreter may plausibly be taken to be able to identify before he can interpret, since he may know that a person intends to express a truth in uttering a sentence without having any idea *what* truth. Not that sincere assertion is the only reason to suppose that a person holds a sentence to be true. Lies, commands, stories, irony, if they are detected as attitudes, can reveal whether a speaker holds his sentences to be true. There is no reason to rule out other attitudes towards sentences, such as wishing true, wanting to make true, believing one is going to make true, and so on, but I am inclined to think that all evidence of this kind may be summed up in terms of holding sentences to be true.

Suppose, then, that the evidence available is just that speakers of the language to be interpreted hold various sentences to be true at certain times and under specified circumstances. How can this evidence be used to support a theory of truth? On the one hand, we have T-sentences, in the form:

(T) 'Es regnet' is true-in-German when spoken by *x* at time *t* if and only if it is raining near *x* at *t*.

On the other hand, we have the evidence, in the form:

(E) Kurt belongs to the German speech community and Kurt holds true 'Es regnet' on Saturday at noon and it is raining near Kurt on Saturday at noon.

We should, I think, consider (E) as evidence that (T) is true. Since (T) is a universally quantified conditional, the first step would be to gather more evidence to support the claim that:

(GE) $(x)(t)$ (if *x* belongs to the German speech community then (*x* holds true 'Es regnet' at *t* if and only if it is raining near *x* at *t*)).

The appeal to a speech community cuts a corner but begs no question: speakers belong to the same speech community if the same theories of interpretation work for them.

The obvious objection is that Kurt, or anyone else, may be wrong about whether it is raining near him. And this is of course a reason for not taking (E) as conclusive evidence for (GE) or for (T); and a reason not to expect generalizations like (GE) to be more than generally true. The method is rather one of getting a best fit. We want a theory that satisfies the formal constraints on a theory of truth, and that maximizes agreement, in the sense of making Kurt (and others) right, as far as we can tell, as often as possible. The concept of maximization cannot be taken literally here, since sentences are infinite in number, and anyway once the theory begins to take shape it makes sense to accept intelligible error and to make allowance for the relative likelihood of various kinds of mistake.[15]

The process of devising a theory of truth for an unknown native tongue might in crude outline go as follows. First we look for the best way to fit our logic, to the extent required to get a theory satisfying Convention T, on to the new language; this may mean reading the logical structure of first-order quantification theory (plus identity) into the language, not taking the logical constants one by one, but treating this much of logic as a grid to be fitted on to the language in one fell swoop. The evidence here is classes of sentences always held true or always held false by almost everyone almost all of the time (potential logical truths) and patterns of inference. The first step identifies predicates, singular terms, quantifiers, connectives, and identity; in theory, it settles matters of logical form. The second step concentrates on sentences with indexicals; those sentences sometimes held true and sometimes false according to discoverable changes in the world. This step in conjunction with the first limits the possibilities for interpreting individual predicates. The last step deals with the remaining sentences, those on which there is not uniform agreement, or whose held truth value does not depend systematically on changes in the environment.[16]

[15] For more on getting a 'best fit' see Essays 10–12.

[16] Readers who appreciate the extent to which this account parallels Quine's account of radical translation in Chapter 2 of *Word and Object* will also notice the differences: the semantic constraint in my method forces quantificational structure on the language to be interpreted, which probably does not leave room for indeterminacy of logical form; the notion of stimulus meaning plays no role in my method, but its place is taken by reference to the objective features of the world which alter in conjunction with changes in attitude towards the truth of sentences; the principle of charity, which Quine emphasizes only in connection with the identification of the (pure) sentential connectives, I apply across the board.

This method is intended to solve the problem of the interdependence of belief and meaning by holding belief constant as far as possible while solving for meaning. This is accomplished by assigning truth conditions to alien sentences that make native speakers right when plausibly possible, according, of course, to our own view of what is right. What justifies the procedure is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement. Applied to language, this principle reads: the more sentences we conspire to accept or reject (whether or not through a medium of interpretation), the better we understand the rest, whether or not we agree about them.

The methodological advice to interpret in a way that optimizes agreement should not be conceived as resting on a charitable assumption about human intelligence that might turn out to be false. If we cannot find a way to interpret the utterances and other behaviour of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything.

Here I would like to insert a remark about the methodology of my proposal. In philosophy we are used to definitions, analyses, reductions. Typically these are intended to carry us from concepts better understood, or clear, or more basic epistemologically or ontologically, to others we want to understand. The method I have suggested fits none of these categories. I have proposed a looser relation between concepts to be illuminated and the relatively more basic. At the centre stands a formal theory, a theory of truth, which imposes a complex structure on sentences containing the primitive notions of truth and satisfaction. These notions are given application by the form of the theory and the nature of the evidence. The result is a partially interpreted theory. The advantage of the method lies not in its free-style appeal to the notion of evidential support but in the idea of a powerful theory interpreted at the most advantageous point. This allows us to reconcile the need for a semantically articulated structure with a theory testable only at the sentential level. The more subtle gain is that very thin evidence in support of each of a potential infinity of points can yield rich results, even with respect to the points. By knowing only the conditions under which speakers hold sentences true, we can come out, given a satisfactory theory, with an interpretation of each sentence. It remains to make good on this last claim. The theory itself at best gives truth conditions. What we need to

show is that if such a theory satisfies the constraints we have specified, it may be used to yield interpretations.

3. *If we know that a theory of truth satisfies the formal and empirical criteria described, can we interpret utterances of the language for which it is a theory?*

A theory of truth entails a T-sentence for each sentence of the object language, and a T-sentence gives truth conditions. It is tempting, therefore, simply to say that a T-sentence 'gives the meaning' of a sentence. Not, of course, by naming or describing an entity that is a meaning, but simply by saying under what conditions an utterance of the sentence is true.

But on reflection it is clear that a T-sentence does not give the meaning of the sentence it concerns: the T-sentences does fix the truth value relative to certain conditions, but it does not say the object language sentence is true *because* the conditions hold. Yet if truth values were all that mattered, the T-sentence for 'Snow is white' could as well say that it is true if and only if grass is green or $2 + 2 = 4$ as say that it is true if and only if snow is white. We may be confident, perhaps, that no satisfactory theory of truth will produce such anomalous T-sentences, but this confidence does not license us to make more of T-sentences.

A move that might seem helpful is to claim that it is not the T-sentence alone, but the canonical proof of a T-sentence, that permits us to interpret the alien sentence. A canonical proof, given a theory of truth, is easy to construct, moving as it does through a string of biconditionals, and requiring for uniqueness only occasional decisions to govern left and right precedence. The proof does reflect the logical form the theory assigns to the sentence, and so might be thought to reveal something about meaning. But in fact we would know no more than before about how to interpret if all we knew was that a certain sequence of sentences was the proof, from some true theory, of a particular T-sentence.

A final suggestion along these lines is that we can interpret a particular sentence provided we know a correct theory of truth that deals with the language of the sentence. For then we know not only the T-sentence for the sentence to be interpreted, but we also 'know' the T-sentences for all other sentences; and of course, all the proofs.

Then we would see the place of the sentence in the language as a whole, we would know the role of each significant part of the sentence, and we would know about the logical connections between this sentence and others.

If we knew that a T-sentence satisfied Tarski's Convention T, we would know that it was true, and we could use it to interpret a sentence because we would know that the right branch of the biconditional translated the sentence to be interpreted. Our present trouble springs from the fact that in radical interpretation we cannot assume that a T-sentence satisfies the translation criterion. What we have been overlooking, however, is that we have supplied an alternative criterion: this criterion is that the totality of T-sentences should (in the sense described above) optimally fit evidence about sentences held true by native speakers. The present idea is that what Tarski assumed outright for each T-sentence can be indirectly elicited by a holistic constraint. If that constraint is adequate, each T-sentence will in fact yield an acceptable interpretation.

A T-sentence of an empirical theory of truth can be used to interpret a sentence, then, provided we also know the theory that entails it, and know that it is a theory that meets the formal and empirical criteria.[17] For if the constraints are adequate, the range of acceptable theories will be such that any of them yields some correct interpretation for each potential utterance. To see how it might work, accept for a moment the absurd hypothesis that the constraints narrow down the possible theories to one, and this one implies the T-sentence (T) discussed previously. Then we are justified in using this T-sentence to interpret Kurt's utterance of 'Es regnet' as his saying that it is raining. It is not likely, given the flexible nature of the constraints, that all acceptable theories will be identical. When all the evidence is in, there will remain, as Quine has emphasized, the trade-offs between the beliefs we attribute to a speaker and the interpretations we give his words. But the resulting indeterminacy cannot be so great but that any theory that passes the tests will serve to yield interpretations.

[17] See footnote 11 of Essay 2 and Essay 12.

§ 11. INTRASUBJECTIVE SYNONYMY OF OCCASION SENTENCES

Stimulus meaning remains defined without regard to observationality. But when applied to non-observational sentences like 'Bachelor' it bears little resemblance to what might reasonably be called meaning. Translation of 'Soltero' as 'Bachelor' manifestly cannot be predicated on identity of stimulus meanings between speakers; nor can synonymy of 'Bachelor' and 'Unmarried man'.

But curiously enough the stimulus meanings of 'Bachelor' and 'Unmarried man' are, despite all this, identical for any one speaker.[1] An individual would at any one time be prompted by the same stimulations to assent to 'Bachelor' and 'Unmarried man'; and similarly for dissent. *Stimulus synonymy,* or sameness of stimulus meaning, is as good a standard of synonymy for non-observational occasion sentences as for observation sentences as long as we stick to one speaker. For each speaker, 'Bachelor' and 'Unmarried man' are stimulus-synonymous without having the same meaning in any acceptably defined sense of 'meaning' (for stimulus meaning is, in the case of 'Bachelor', nothing of the kind). Very well; here is a case where we may welcome the synonymy and let the meaning go.

The one-speaker restriction presents no obstacle to saying that 'Bachelor' and 'Unmarried man' are stimulus-synonymous for the whole community, in the sense of being thus for each member. A

[1] It can be argued that this much-used example of synonymy has certain imperfections having to do with ages, divorce, and bachelors of arts. Another example much used in philosophy, 'brother' and 'male sibling', may be held to bog down under certain church usages. An example that is perhaps unassailable is 'mother's father' and 'maternal grandfather' (poetic connotations not being here in point), or 'widower' and 'man who lost his wife' (Jakobson). However, with this much by way of caveat against quibbling, perhaps we can keep to our conventional example and overlook its divagations.

QUINE, W.V.O.
WORD AND OBJECT

practical extension even to the two-language case is not far to seek if a bilingual speaker is at hand. 'Bachelor' and 'Soltero' will be stimulus-synonymous for him. Taking him as a sample, we may treat 'Bachelor' and 'Soltero' as synonymous for the translation purposes of the two whole linguistic communities that he represents. Whether he is a good enough sample would be checked by observing the fluency of his communication in both communities and by comparing other bilinguals.

Section 10 left the linguist unable to guess the trend of the stimulus meaning of a non-observational occasion sentence from sample cases. We now see a way, though costly, in which he can still accomplish radical translation of such sentences. He can settle down and learn the native language directly as an infant might.[2] Having thus become bilingual, he can translate the non-observational occasion sentences by introspected stimulus synonymy.

This step has the notable effect of initiating clear recognition of native falsehoods. As long as the linguist does no more than correlate the native's observation sentences with his own by stimulus meaning, he cannot discount any of the native's verdicts as false—unless *ad hoc,* most restrainedly, to simplify his correlations. But once he becomes bilingual and so transcends the observation sentences, he can bicker with the native as a brother.

Even short of going bilingual there is no difficulty in comparing two non-observational native sentences to see if they are intrasubjectively stimulus-synonymous for the native. The linguist can do this without having intuitively conjectured the trend of stimulus meaning of either sentence. He need merely query the sentences in parallel under random stimulations until he either hits a stimulation that prompts assent or dissent to one sentence and not to the other, or else is satisfied at last that he is not going to. A visiting Martian who never learns under what circumstances to apply 'Bachelor', or 'Unmarried man' either, can still find out by the above method that 'Bachelor' for one English speaker does not have the same stimulus meaning as 'Bachelor' for a different English speaker and that it has the same as 'Unmarried man' for the same speaker. He can, anyway, apart from one difficulty: there is no evident reason why it should occur to him thus blindly to try comparing 'Unmarried man' with 'Bachelor'. This difficulty makes the intrasubjec-

[2] See Chapter III for reflections on the infant's learning of our own language.

tive stimulus synonymy of non-observational occasion sentences less readily accessible to an alien linguist than the stimulus synonymy of observation sentences such as 'Gavagai' and 'Rabbit'. Still the linguist can examine for intrasubjective stimulus synonymy any pair of native occasion sentences that it occurs to him to wonder about; and we shall see in § 15 how indirect considerations can even suggest such pairs for examination.

Between the stimulus meaning of any sentence for one man and the stimulus meaning of the same or any other sentence for another man there are almost bound to be countless discrepancies in point of verbally contaminated stimulations, as long as one man understands a language that the other does not. The argument is that of the kibitzer case in § 9. The translating linguist had for this reason to discount verbally contaminated discrepancies. But intrasubjective comparisons are free of this trouble. Intrasubjectively we can even compare the occasion sentences 'Yes', 'Uh huh', and 'Quite' for stimulus synonymy, though the stimulations that enter into the stimulus meanings of these sentences are purely verbal in their relevant portions. A further advantage of the intrasubjective situation appears in the case of stimulations that would at a given time shock one speaker and not another into silence (cf. § 9); for clearly these will constitute no discrepancies intrasubjectively. Altogether the equating of stimulus meanings works out far better intrasubjectively than between subjects: it goes beyond observation sentences, it absorbs shock, and it better accommodates verbal stimulations.

Verbal stimulations can plague even the intrasubjective comparisons when they are stimulations of "second intention"—i.e., when besides consisting of words they are about words. Second-intention examples are the bane of theoretical linguistics, also apart from synonymy studies. Thus take the linguist engaged in distinguishing between those sequences of sounds or phonemes that can occur in English speech and those that cannot: all his excluded forms can return to confound him in second-intention English, as between quotation marks. Now some second-intention stimulations that could prompt a subject to assent to one of the queries 'Bachelor?' and 'Unmarried man?' to the exclusion of the other are as follows: a stimulation presenting the spelling of 'bachelor'; a stimulation presenting the words 'rhymes with 'harried man''; a stimulation presenting a glimpse of a bachelor friend together with a plea to

redefine 'bachelor'. It is not easy to find a behavioral criterion of second-intention whereby to screen such cases, especially the last.

Leaving that problem unsolved, we have still to note another and more humdrum restriction that needs to be observed in equating sentences by stimulus meanings: we should stick to short sentences. Otherwise subjects' mere incapacity to digest long questions can, under our definitions, issue in difference of stimulus meanings between long and short sentences which we should prefer to find synonymous. A stimulation may prompt assent to the short sentence and not to the long one just because of the opacity of the long one; yet we should then like to say not that the subject has shown the meaning of the long sentence to be different, but merely that he has failed to encompass it. Still a concept of synonymy initially significant only for short sentences can be extended to long sentences by analogy, e.g. as follows. By a *construction*, linguistically speaking, let us understand any fixed way of building a composite expression from arbitrary components of appropriate sort, one or more at a time. (What is fixed may include certain additive words, as well as the way of arranging the unfixed components.) Now two sentence-forming constructions may be so related that whenever applied to the same components they yield mutually synonymous results, as long as the results are short enough to be compared for synonymy. In this event it is natural, by extension, to count also as mutually synonymous any results of applying those constructions to identical components however long. But to simplify ensuing considerations let us continue to reason without reference to this refinement where we can.

Our success with 'Bachelor' and 'Unmarried man' has been sufficient, despite the impasse at second intention, to tempt us to overestimate how well intrasubjective stimulus synonymy withstands collateral information. By way of corrective, consider the Himalayan explorer who has learned to apply 'Everest' to a distant mountain seen from Tibet and 'Gaurisanker' to one seen from Nepal. As occasion sentences these words have mutually exclusive stimulus meanings for him until his explorations reveal, to the surprise of all concerned, that the peaks are identical. His discovery is painfully empirical, not lexicographic; nevertheless the stimulus meanings of 'Everest' and 'Gaurisanker' coincide for him thenceforward.[8]

[8] I am indebted to Davidson for this point and to Schrödinger, *What Is Life?*, for the example. I am told that the example is wrong geographically.

Or again consider the occasion sentences 'Indian nickel' and 'Buffalo nickel'. These have distinct stimulus meanings for a boy for his first minute or two of passive acquaintance with these coins, and when he gets to turning them over the stimulus meanings tend to fuse.

Do they fully fuse? The question whether 'Indian nickel' and 'Buffalo nickel' have the same stimulus meaning for a given subject is the question whether any sequence of ocular irradiations or other stimulation (within the modulus), realized or not, *would* now prompt the subject to assent to or dissent from 'Indian nickel' and not 'Buffalo nickel' or vice versa. Among such stimulations are those that present, to all appearances, a coin whose obverse is like that of an Indian nickel but whose reverse bears some device other than the buffalo. Such stimulations can with a little felony even be realized. After a modulus-long examination of such a hybrid coin, a novice might conclude with surprise that there are after all two kinds of Indian nickel, while an expert, sure of his numismatics, might conclude that the coin must be fraudulent. For the expert, 'Indian nickel' and 'Buffalo nickel' are stimulus-synonymous; for the novice not.

The novice does believe and continues to believe, as the expert does, that all Indian nickels are buffalo nickels and vice versa; for the novice has not been and will not be actually subjected to the surprising stimulation described. But the mere fact that there is such a stimulation pattern and that the novice *would* now thus respond to it (whether we know it or not) is what, by definition, makes the stimulus meanings of 'Indian nickel' and 'Buffalo nickel' differ for the novice even as of now.

To keep our example pertinent we must abstract from what may be called the conniving mode of speech: the mode in which we knowingly speak of Olivier as Macbeth, of a statue of a horse as a horse, of a false nickel as a nickel. Even the expert would in practice speak of the prepared coin as "that Indian nickel with the whoozis on the back," adding that it was phony. Here we have a broader usage of 'nickel', under which nobody would seriously maintain even that all Indian nickels are in point of fact buffalo nickels and vice versa; whereas our purpose in the example is to examine two supposedly coextensive terms for sameness of stimulus meaning. In the example, therefore, read 'Indian nickel' and 'buffalo nickel' as 'real Indian nickel', 'real buffalo nickel'.

From the example we see that two terms can in fact be coextensive, or true of the same things, without being intrasubjectively stimulus-synonymous as occasion sentences. They can be believed coextensive without being, even for the believer, stimulus-synonymous as occasion sentences; witness 'Indian nickel' and 'Buffalo nickel' for the novice. But when as in the expert's case the belief is so firm that no pattern of stimulation (within the modulus) would suffice to dislodge it, they are stimulus-synonymous as occasion sentences.

So it is apparent that intrasubjective stimulus synonymy remains open to criticism, from intuitive preconceptions, for relating occasion sentences whose stimulus meanings coincide on account of collateral information. Now there is still a way of cutting out the effects of idiosyncratic information: we can hold out for virtual constancy over the community. In this social sense of stimulus synonymy, 'Indian nickel' and 'Buffalo nickel' would cease to count as stimulus-synonymous, because of such speakers as our novice; whereas 'Bachelors' and 'Unmarried man' might still rate as stimulus-synonymous even socially, as being intrasubjectively stimulus-synonymous for nearly everybody. There is still no screen against the effects of collateral information common to the community; but, as urged in § 9, I think that at that point the ideal becomes illusory.

ON THE REASONS FOR INDETERMINACY OF TRANSLATION *

MY *gavagai* example has figured too centrally in discussions of the indeterminacy of translation. Readers see the example as the ground of the doctrine, and hope by resolving the example to cast doubt on the doctrine. The real ground of the doctrine is very different, broader and deeper.

Let us put translation aside for a while and think about physical theory. Naturally it is underdetermined by past evidence; a

* I am indebted to Burton Dreben for helpful criticism of an earlier draft of this paper.

QUINE, W.V.O.

IN: JOURNAL OF PHILOSOPHY 67(1970)

future observation can conflict with it. Naturally it is underdetermined by past and future evidence combined, since some observable event that conflicts with it can happen to go unobserved. Moreover many people will agree, far beyond all this, that physical theory is underdetermined even by all *possible* observations. Not to make a mystery of this mode of possibility, what I mean is the following. Consider all the observation sentences of the language: all the occasion sentences that are suited for use in reporting observable events in the external world.[1] Apply dates and positions to them in all combinations, without regard to whether observers were at the place and time. Some of these place-timed sentences will be true and the others false, by virtue simply of the observable though unobserved past and future events in the world. Now my point about physical theory is that physical theory is underdetermined even by all these truths. Theory can still vary though all possible observations be fixed. Physical theories can be at odds with each other and yet compatible with all possible data even in the broadest sense. In a word, they can be logically incompatible and empirically equivalent. This is a point on which I expect wide agreement, if only because the observational criteria of theoretical terms are commonly so flexible and fragmentary. People who agree on this general point need not agree as to how much of physical theory is empirically unfixed in this strong sense; some will acknowledge such slack only in the highest and most speculative reaches of physical theory, while others see it as extending even to common-sense traits of macroscopic bodies.

Now let us turn to the radical translation of a radically foreign physicist's theory. As always in radical translation, the starting point is the equating of observation sentences of the two languages by an inductive equating of stimulus meanings. In order afterward to construe the foreigner's theoretical sentences we have to project analytical hypotheses, whose ultimate justification is substantially just that the implied observation sentences match up. But now the same old empirical slack, the old indeterminacy between physical theories, recurs in second intension. Insofar as the truth of a physical theory is underdetermined by observables, the translation of the foreigner's physical theory is underdetermined by translation of his observation sentences. If our physical theory can vary though all possible observations be fixed, then our translation of his physical theory can

[1] The concept of observation sentence that I developed in §10 of *Word and Object* gains perhaps some further clarification in pp. 85–89 of *Ontological Relativity and Other Essays* (New York: Columbia University Press, 1969).

vary though our translations of all possible observation reports on his part be fixed. Our translation of his observation sentences no more fixes our translation of his physical theory than our own possible observations fix our own physical theory.

The indeterminacy of translation is not just an instance of the empirically underdetermined character of physics. The point is not just that linguistics, being a part of behavioral science and hence ultimately of physics, shares the empirically underdetermined character of physics. On the contrary, the indeterminacy of translation is additional. Where physical theories *A* and *B* are both compatible with all possible data, we might adopt *A* for ourselves and still remain free to translate the foreigner either as believing *A* or as believing *B*.

Such choice between *A* and *B* in translation could be guided by simplicity. By imputing *B* to the foreigner we might come out with shorter and more direct translations, and with less in the way of elaborate contextual paraphrases, than by imputing *A* to him. That is one possibility. A second possibility is that both choices, *A* and *B*, require forbiddingly circuitous and cumbersome translation rules. In this case we might regard the foreigner as holding neither *A* nor *B*; we might attribute to him rather some false physical theory which we can refute, or some obscure one which we despair of penetrating, or we might even regard him as holding no coherent physical theory at all. But we can imagine also, third, the possibility that *A* and *B* are both reasonably attributable. It might turn out that with just moderate circuitousness of translation at certain points —different points—*A* and *B* could be imputed about equally well. In this event no basis for a choice can be gained by exposing the foreigner to new physical data and noting his verbal response, since the theories *A* and *B* fit all possible observations equally well. No basis can be gained by interrogation in a theoretical vein, since the interrogation would take place in the foreigner's language and so could itself be interpreted according to either plan. In this event our choice would be determined simply by the accident of hitting upon one of the two systems of translation first.

The metaphor of the black box, often so useful, can be misleading here. The problem is not one of hidden facts, such as might be uncovered by learning more about the brain physiology of thought processes. To expect a distinctive physical mechanism behind every genuinely distinct mental state is one thing; to expect a distinctive mechanism for every purported distinction that can be phrased in traditional mentalistic language is another. The question whether,

in the situation last described, the foreigner *really* believes *A* or believes rather *B*, is a question whose very significance I would put in doubt. This is what I am getting at in arguing the indeterminacy of translation.

My argument in these pages has been and will remain directed to you who already agree that there can be logically incompatible and empirically equivalent physical theories *A* and *B*. What degree of indeterminacy of translation you must then recognize, granted the force of my argument, will depend on the amount of empirical slack that you are willing to acknowledge in physics. If you were one of those who saw physics as empirically underdetermined only in its highest theoretical reaches, then by the argument at hand I can claim your concurrence in the indeterminacy of translation only of highly theoretical physics. For my own part, I think the empirical slack in physics extends to ordinary traits of ordinary bodies and hence that the indeterminacy of translation likewise affects that level of discourse. But it is important, for those who would not go so far, to note the graduation of liabilities.

Gavagai, whose troubles I shall now review, lay at an extreme of the scale. It was an observation sentence. Its stimulus meaning was inductively well established, we supposed, coinciding with that of 'Rabbit'.[2] Where indeterminacy threatened was in trying to settle upon the divided reference of *gavagai* as a term: whether rabbits or rabbit stages or undetached rabbit parts. Readers have responded with suggestions of how, with help of screens or other devices, we might hope to give the native informant an inkling of the desired distinctions and so settle the reference.

Ingenuity in this vein proves unrewarding because of vagueness of purpose. The purpose cannot be to drive a wedge between stimulus meanings of observation sentences, thereby linking *Gavagai* rather to 'Rabbit' than to 'Rabbit stage' or 'Undetached rabbit part'; for the stimulus meanings of all these sentences are incontestably identical. They comprise the stimulations that would make people think a rabbit was present. The purpose can only be to settle what *gavagai* denotes for the native as a term. But the whole notion of terms and their denotation is bound up with our own grammatical analysis of the sentences of our own language. It can

[2] Strictly speaking, even this induction presupposes something like an analytical hypothesis in a small way: the decision as to what to take as signs of assent and dissent. See *Word and Object*, p. 30; also D. Davidson and J. Hintikka, eds., *Words and Objections* (Dordrecht: Reidel, 1968), pp. 312, 317, or *Synthèse*, XIX, 1/2 (December 1968): 284, 289.

be projected on the native language only as we settle what to count in the native language as analogues of our pronouns, identity, plurals, and related apparatus; and I urged in *Word and Object* that there would be some freedom of choice on this score. Once such choices are settled, on the other hand, however arbitrarily, the question whether the *gavagai* are rabbits or stages or parts can be settled too, by interrogation.

The most to hope for from the screens and kindred aids, then, is an indirect hint as to which of various analytical hypotheses regarding pronouns, identity, plurals, etc. might in the end work out most naturally. When this kind of hint is available, should we say that the supposed multiplicity of choices was not in fact open after all? Or should we say that the choice is open but that we have found a practical consideration that will help us in choosing? The issue is palpably unreal, and the doctrine of the indeterminacy of translation depends in no way upon it.

The *gavagai* example was at best an example only of the inscrutability of terms, not of the indeterminacy of translation of sentences. As sentence, *Gavagai* had a translation that was unique to within stimulus synonymy; for the occasion sentences 'Rabbit', 'Rabbit stage', and 'Undetached rabbit part' are stimulus-synonymous and holophrastically interchangeable. The *gavagai* example had only this indirect bearing on indeterminacy of translation of sentences: one could imagine with some plausibility that some lengthy nonobservational sentences containing *gavagai* could be found which would go into English in materially different ways according as *gavagai* was equated with one or another of the terms 'rabbit', 'rabbit stage', etc. This whole effort was aimed not at proof but at helping the reader to reconcile the indeterminacy of translation imaginatively with the concrete reality of radical translation. The argument for the indeterminacy is another thing, as seen earlier in this paper.

Over the inscrutability of terms itself there is little room for debate. A clear example from real life was seen in connection with the Japanese classifiers.[3] This example makes it pretty clear, moreover, that the inscrutability of terms need not always bring indeterminacy of sentence translation in its train, however the case may be in particular with *gavagai*. Again the questions raised by deferred ostension (*op. cit.*), e.g., as between expressions and their Gödel numbers, are strictly a matter of inscrutability of terms.

[3] *Ontological Relativity and Other Essays*, pp. 35ff. Also in this JOURNAL, LXV, 7 (April 4, 1968): 191ff.

This, not the indeterminacy of translation, is the substance of ontological relativity.

There are two ways of pressing the doctrine of indeterminacy of translation to maximize its scope. I can press from above and press from below, playing both ends against the middle. At the upper end there is the argument, early in the present paper, which is meant to persuade anyone to recognize the indeterminacy of translation of such portions of natural science as he is willing to regard as underdetermined by all possible observations. If I can get people to see this empirical slack as affecting not just highly theoretical physics but fairly common-sense talk of bodies, then I can get them to concede indeterminacy of translation of fairly common-sense talk of bodies. This I call pressing from above.

By pressing from below I mean pressing whatever arguments for indeterminacy of translation can be based on the inscrutability of terms. I suppose Harman's example[4] regarding natural numbers comes under this head, theoretical though it is. It is that the sentence '3 ϵ 5' goes into a true sentence of set theory under von Neumann's way of construing natural numbers, but goes into a false one under Zermelo's way. But a limitation of this example, as Harman recognizes, is that '3 ϵ 5' rates as nonsense apart from set-theoretic explications of natural number.

In these pages I prefer not to speculate on how much better one might do from below, or from above either. My purpose here is to separate the issues and identify the arguments; and this may be managed most effectively by leaving the reader to consider what more might be proved.

W. V. QUINE
Harvard University

AUSTIN, JOHN

HOW TO DO THINGS
WITH WORDS

LECTURES I, V, VI, VIII, XI

LECTURE I

WHAT I shall have to say here is neither difficult nor contentious; the only merit I should like to claim for it is that of being true, at least in parts. The phenomenon to be discussed is very widespread and obvious, and it cannot fail to have been already noticed, at least here and there, by others. Yet I have not found attention paid to it specifically.

It was for too long the assumption of philosophers that the business of a 'statement' can only be to 'describe' some state of affairs, or to 'state some fact', which it must do either truly or falsely. Grammarians, indeed, have regularly pointed out that not all 'sentences' are (used in making) statements:[1] there are, traditionally, besides (grammarians') statements, also questions and exclamations, and sentences expressing commands or wishes or concessions. And doubtless philosophers have not intended to deny this, despite some loose use of 'sentence' for 'statement'. Doubtless, too, both grammarians and philosophers have been aware that it is by no means easy to distinguish even questions, commands, and so on from statements by means of the few and jejune grammatical marks available, such as word order, mood, and the like:

[1] It is, of course, not really correct that a sentence ever *is* a statement: rather, it is *used* in *making a statement*, and the statement itself is a 'logical construction' out of the makings of statements.

though perhaps it has not been usual to dwell on the difficulties which this fact obviously raises. For how do we decide which is which? What are the limits and definitions of each?

But now in recent years, many things which would once have been accepted without question as 'statements' by both philosophers and grammarians have been scrutinized with new care. This scrutiny arose somewhat indirectly—at least in philosophy. First came the view, not always formulated without unfortunate dogmatism, that a statement (of fact) ought to be 'verifiable', and this led to the view that many 'statements' are only what may be called pseudo-statements. First and most obviously, many 'statements' were shown to be, as KANT perhaps first argued systematically, strictly nonsense, despite an unexceptionable grammatical form: and the continual discovery of fresh types of nonsense, unsystematic though their classification and mysterious though their explanation is too often allowed to remain, has done on the whole nothing but good. Yet we, that is, even philosophers, set some limits to the amount of nonsense that we are prepared to admit we talk: so that it was natural to go on to ask, as a second stage, whether many apparent pseudo-statements really set out to be 'statements' at all. It has come to be commonly held that many utterances which look like statements are either not intended at all, or only intended in part, to record or impart straightforward information about the facts: for example, 'ethical propositions' are perhaps intended, solely or partly, to evince

emotion or to prescribe conduct or to influence it in special ways. Here too KANT was among the pioneers. We very often also use utterances in ways beyond the scope at least of traditional grammar. It has come to be seen that many specially perplexing words embedded in apparently descriptive statements do not serve to indicate some specially odd additional feature in the reality reported, but to indicate (not to report) the circumstances in which the statement is made or reservations to which it is subject or the way in which it is to be taken and the like. To overlook these possibilities in the way once common is called the 'descriptive' fallacy; but perhaps this is not a good name, as 'descriptive' itself is special. Not all true or false statements are descriptions, and for this reason I prefer to use the word 'Constative'. Along these lines it has by now been shown piecemeal, or at least made to look likely, that many traditional philosophical perplexities have arisen through a mistake—the mistake of taking as straightforward statements of fact utterances which are *either* (in interesting non-grammatical ways) nonsensical *or else* intended as something quite different.

Whatever we may think of any particular one of these views and suggestions, and however much we may deplore the initial confusion into which philosophical doctrine and method have been plunged, it cannot be doubted that they are producing a revolution in philosophy. If anyone wishes to call it the greatest and most salutary in its history, this is not, if you come to think of it, a

large claim. It is not surprising that beginnings have been piecemeal, with *parti pris*, and for extraneous aims; this is common with revolutions.

PRELIMINARY ISOLATION OF THE PERFORMATIVE[1]

The type of utterance we are to consider here is not, of course, in general a type of nonsense; though misuse of it can, as we shall see, engender rather special varieties of 'nonsense'. Rather, it is one of our second class—the masqueraders. But it does not by any means necessarily masquerade as a statement of fact, descriptive or constative. Yet it does quite commonly do so, and that, oddly enough, when it assumes its most explicit form. Grammarians have not, I believe, seen through this 'disguise', and philosophers only at best incidentally.[2] It will be convenient, therefore, to study it first in this misleading form, in order to bring out its characteristics by contrasting them with those of the statement of fact which it apes.

We shall take, then, for our first examples some utterances which can fall into no hitherto recognized *grammatical* category save that of 'statement', which are not nonsense, and which contain none of those verbal danger-signals which philosophers have by now detected or think

[1] Everything said in these sections is provisional, and subject to revision in the light of later sections.

[2] Of all people, jurists should be best aware of the true state of affairs. Perhaps some now are. Yet they will succumb to their own timorous fiction, that a statement of 'the law' is a statement of fact.

they have detected (curious words like 'good' or 'all', suspect auxiliaries like 'ought' or 'can', and dubious constructions like the hypothetical): all will have, as it happens, humdrum verbs in the first person singular present indicative active.[1] Utterances can be found, satisfying these conditions, yet such that

A. they do not 'describe' or 'report' or constate anything at all, are not 'true or false'; and
B. the uttering of the sentence is, or is a part of, the doing of an action, which again would not *normally* be described as, or as 'just', saying something.

This is far from being as paradoxical as it may sound or as I have meanly been trying to make it sound: indeed, the examples now to be given will be disappointing.

Examples:

(E. *a*) 'I do (sc. take this woman to be my lawful wedded wife)'—as uttered in the course of the marriage ceremony.[2]

(E. *b*) 'I name this ship the *Queen Elizabeth*'—as uttered when smashing the bottle against the stem.

(E. *c*) 'I give and bequeath my watch to my brother'—as occurring in a will.

(E. *d*) 'I bet you sixpence it will rain tomorrow.'

[1] Not without design: they are all 'explicit' performatives, and of that prepotent class later called 'exercitives'.

[2] [Austin realized that the expression 'I do' is not used in the marriage ceremony too late to correct his mistake. We have let it remain in the text as it is philosophically unimportant that it is a mistake. J. O. U.]

In these examples it seems clear that to utter the sentence (in, of course, the appropriate circumstances) is not to *describe* my doing of what I should be said in so uttering to be doing[1] or to state that I am doing it: it is to do it. None of the utterances cited is either true or false: I assert this as obvious and do not argue it. It needs argument no more than that 'damn' is not true or false: it may be that the utterance 'serves to inform you'—but that is quite different. To name the ship *is* to say (in the appropriate circumstances) the words 'I name, &c.'. When I say, before the registrar or altar, &c., 'I do', I am not reporting on a marriage: I am indulging in it.

What are we to call a sentence or an utterance of this type?[2] I propose to call it a *performative sentence* or a performative utterance, or, for short, 'a performative'. The term 'performative' will be used in a variety of cognate ways and constructions, much as the term 'imperative' is.[3] The name is derived, of course, from 'perform', the usual verb with the noun 'action': it indicates that the issuing of the utterance is the performing of an action

[1] Still less anything that I have already done or have yet to do.

[2] 'Sentences' form a class of 'utterances', which class is to be defined, so far as I am concerned, grammatically, though I doubt if the definition has yet been given satisfactorily. With performative utterances are contrasted, for example and essentially, 'constative' utterances: to issue a constative utterance (i.e. to utter it with a historical reference) is to make a statement. To issue a performative utterance is, for example, to make a bet. See further below on 'illocutions'.

[3] Formerly I used 'performatory': but 'performative' is to be preferred as shorter, less ugly, more tractable, and more traditional in formation.

—it is not normally thought of as just saying something.

A number of other terms may suggest themselves, each of which would suitably cover this or that wider or narrower class of performatives: for example, many performatives are *contractual* ('I bet') or *declaratory* ('I declare war') utterances. But no term in current use that I know of is nearly wide enough to cover them all. One technical term that comes nearest to what we need is perhaps 'operative', as it is used strictly by lawyers in referring to that part, i.e. those clauses, of an instrument which serves to effect the transaction (conveyance or what not) which is its main object, whereas the rest of the document merely 'recites' the circumstances in which the transaction is to be effected.[1] But 'operative' has other meanings, and indeed is often used nowadays to mean little more than 'important'. I have preferred a new word, to which, though its etymology is not irrelevant, we shall perhaps not be so ready to attach some preconceived meaning.

CAN SAYING MAKE IT SO?

Are we then to say things like this:

'To marry is to say a few words', or

'Betting is simply saying something'?

Such a doctrine sounds odd or even flippant at first, but with sufficient safeguards it may become not odd at all.

[1] I owe this observation to Professor H. L. A. Hart.

A sound initial objection to them may be this; and it is not without some importance. In very many cases it is possible to perform an act of exactly the same kind *not* by uttering words, whether written or spoken, but in some other way. For example, I may in some places effect marriage by cohabiting, or I may bet with a totalisator machine by putting a coin in a slot. We should then, perhaps, convert the propositions above, and put it that 'to say a few certain words is to marry' or 'to marry is, in some cases, simply to say a few words' or 'simply to say a certain something is to bet'.

But probably the real reason why such remarks sound dangerous lies in another obvious fact, to which we shall have to revert in detail later, which is this. The uttering of the words is, indeed, usually a, or even *the*, leading incident in the performance of the act (of betting or what not), the performance of which is also the object of the utterance, but it is far from being usually, even if it is ever, the *sole* thing necessary if the act is to be deemed to have been performed. Speaking generally, it is always necessary that the *circumstances* in which the words are uttered should be in some way, or ways, *appropriate*, and it is very commonly necessary that either the speaker himself or other persons should *also* perform certain *other* actions, whether 'physical' or 'mental' actions or even acts of uttering further words. Thus, for naming the ship, it is essential that I should be the person appointed to name her, for (Christian) marrying, it is essential that I should not be already married with a wife

living, sane and undivorced, and so on: for a bet to have been made, it is generally necessary for the offer of the bet to have been accepted by a taker (who must have done something, such as to say 'Done'), and it is hardly a gift if I *say* 'I give it you' but never hand it over.

So far, well and good. The action may be performed in ways other than by a performative utterance, and in any case the circumstances, including other actions, must be appropriate. But we may, in objecting, have something totally different, and this time quite mistaken, in mind, especially when we think of some of the more awe-inspiring performatives such as 'I promise to . . . '. Surely the words must be spoken 'seriously' and so as to be taken 'seriously'? This is, though vague, true enough in general—it is an important commonplace in discussing the purport of any utterance whatsoever. I must not be joking, for example, nor writing a poem. But we are apt to have a feeling that their being serious consists in their being uttered as (merely) the outward and visible sign, for convenience or other record or for information, of an inward and spiritual act: from which it is but a short step to go on to believe or to assume without realizing that for many purposes the outward utterance is a description, *true or false*, of the occurrence of the inward performance. The classic expression of this idea is to be found in the *Hippolytus* (l. 612), where Hippolytus says

ἡ γλῶσσ' ὀμώμοχ', ἡ δὲ φρὴν ἀνωμοτός,

i.e. 'my tongue swore to, but my heart (or mind or other

backstage artiste) did not'.[1] Thus 'I promise to . . .' obliges me—puts on record my spiritual assumption of a spiritual shackle.

It is gratifying to observe in this very example how excess of profundity, or rather solemnity, at once paves the way for immodality. For one who says 'promising is not merely a matter of uttering words! It is an inward and spiritual act!' is apt to appear as a solid moralist standing out against a generation of superficial theorizers: we see him as he sees himself, surveying the invisible depths of ethical space, with all the distinction of a specialist in the *sui generis*. Yet he provides Hippolytus with a let-out, the bigamist with an excuse for his 'I do' and the welsher with a defence for his 'I bet'. Accuracy and morality alike are on the side of the plain saying that *our word is our bond*.

If we exclude such fictitious inward acts as this, can we suppose that any of the other things which certainly are normally required to accompany an utterance such as 'I promise that . . .' or 'I do (take this woman . . .)' are in fact described by it, and consequently do by their presence make it true or by their absence make it false? Well, taking the latter first, we shall next consider what we actually do say about the utterance concerned when one or another of its normal concomitants is *absent*. In no case do we say that the utterance was false but rather

[1] But I do not mean to rule out all the offstage performers—the lights men, the stage manager, even the prompter; I am objecting only to certain officious understudies, who would duplicate the play.

that the utterance—or rather the *act*,[1] e.g. the promise—was void, or given in bad faith, or not implemented, or the like. In the particular case of promising, as with many other performatives, it is appropriate that the person uttering the promise should have a certain intention, viz. here to keep his word: and perhaps of all concomitants this looks the most suitable to be that which 'I promise' does describe or record. Do we not actually, when such intention is absent, speak of a 'false' promise? Yet so to speak is *not* to say that the utterance 'I promise that . . .' is false, in the sense that though he states that he does, he doesn't, or that though he describes he misdescribes—misreports. For he *does* promise: the promise here is not even *void*, though it is given *in bad faith*. His utterance is perhaps misleading, probably deceitful and doubtless wrong, but it is not a lie or a misstatement. At most we might make out a case for saying that it implies or insinuates a falsehood or a misstatement (to the effect that he does intend to do something): but that is a very different matter. Moreover, we do not speak of a false bet or a false christening; and that we *do* speak of a false promise need commit us no more than the fact that we speak of a false move. 'False' is not necessarily used of statements only.

[1] We deliberately avoid distinguishing these, precisely because the distinction is not in point.

LECTURE V

AT the end of the previous lecture we were reconsidering the question of the relations between the performative utterance and statements of various kinds which certainly are true or false. We mentioned as specially notable four such connexions:

(1) If the performative utterance 'I apologize' is happy, then the statement that I am apologizing is true.

(2) If the performative utterance 'I apologize' is to be happy, then the statement that certain conditions obtain —those notably in Rules A. 1 and A. 2—must be true.

(3) If the performative utterance 'I apologize' is to be happy, then the statement that certain other conditions obtain—those notably in our rule Γ. 1—must be true.

(4) If performative utterances of at least some kinds are happy, for example contractual ones, then statements typically of the form that I ought or ought not subsequently to do some particular thing are true.

I was saying that there seemed to be some similarity, and perhaps even an identity, between the second of these connexions and the phenomenon which has been called, in the case of statements as opposed to performatives, 'presupposition': and likewise between the third of these connexions and the phenomenon called (sometimes and not, to my mind, correctly) in the case of statements,

'implication'; these, presupposition and implication, being two ways in which the truth of a statement may be connected importantly with the truth of another without it being the case that the one entails the other in the sole sort of sense preferred by obsessional logicians. Only the fourth and last of the above connexions could be made out—I do not say how satisfactorily—to resemble entailment between statements. 'I promise to do *X* but I am under no obligation to do it' may certainly look more like a self-contradiction—whatever that is—than 'I promise to do *X* but I do not intend to do it': also 'I am under no obligation to do *p*' might be held to entail 'I did not promise to do *p*', and one might think that the way in which asserting *p* commits me to asserting *q* is not unlike the way in which promising to do *X* commits me to doing *X*. But I do not want to say that there is or is not any parallel here; only that at least there is a very close parallel in the other two cases; which suggest that at least in some ways there is danger of our initial and tentative distinction between constative and performative utterances breaking down.

We may, however, fortify ourselves in the conviction that the distinction is a final one by reverting to the old idea that the constative utterance is true or false and the performative is happy or unhappy. Contrast the fact that I am apologizing, which depends on the performative 'I apologize' being happy, with the case of the statement 'John is running', which depends for its truth on its being the fact or case that John is running. But perhaps

this contrast is not so sound either: for, to take statements first, connected with the utterance (constative) 'John is running' is the statement 'I am stating that John is running': and this may depend for its truth on the happiness of 'John is running', just as the truth of 'I am apologizing' depends on the happiness of 'I apologize'. And, to take performatives second: connected with the performative (I presume it is one) 'I warn you that the bull is about to charge' is the fact, if it is one, that the bull is about to charge: if the bull is *not*, then indeed the utterance 'I warn you that the bull is about to charge' is open to criticism—but not in any of the ways we have hitherto characterized as varieties of unhappiness. We should not in this case say the warning was void—i.e. that he did not warn but only went through a form of warning—nor that it was insincere: we should feel much more inclined to say the warning was false or (better) mistaken, as with a statement. So that considerations of the happiness and unhappiness type may infect statements (or some statements) and considerations of the type of truth and falsity may infect performatives (or some performatives).

We have then to take a further step out into the desert of comparative precision. We must ask: is there some precise way in which we can definitely distinguish the performative from the constative utterance? And in particular we should naturally ask first whether there is some *grammatical* (or lexicographical) criterion for distinguishing the performative utterance.

So far we have considered only a small number of classic examples of performatives, all with verbs in the first person singular present indicative active. We shall see very shortly that there were good reasons for this piece of slyness. Examples are 'I name', 'I do', 'I bet', 'I give'. There are fairly obvious reasons, with which I shall nevertheless shortly deal, why this is the commonest type of explicit performative. Note that 'present' and 'indicative' are, of course, both misnomers (not to mention the misleading implications of 'active')—I am only using them in the well-known grammatical way. For example the 'present', as distinct from 'continous present', is normally nothing to do with describing (or even indicating) what I am doing at present. 'I drink beer', as distinct from 'I am drinking beer', is not analogous to a future and a past tense describing what I shall do in the future or have done in the past. It is really more commonly the *habitual* indicative, when it is 'indicative' at all. And where it is not habitual but in a way 'present' genuinely, as in a way it is in performatives, if you like, such as 'I name', then it is certainly not 'indicative' in the sense grammarians intend, that is reporting, describing, or informing about an actual state of affairs or occurrent event: because, as we have seen, it does not describe or inform at all, but is used for, or in, the doing of something. So we use 'present indicative' merely to mean the English grammatical form 'I name', 'I run', &c. (This mistake in terminology is due to assimilating, for example, 'I run' to the Latin *curro*, which should really generally be

translated 'I am running'; Latin does not have two tenses where we do.)

Well, is the use of the first person singular and of the present indicative active, so called, essential to a performative utterance? We need not waste our time on the obvious exception of the first person plural, '*we* promise . . .', 'we consent', &c. There are more important and obvious exceptions all over the place (some of which have already been alluded to in passing).

A very common and important type of, one would think, indubitable performative has the verb in the *second or third person* (singular or plural) and the verb in the *passive* voice: so person and voice anyway are not essential. Some examples of this type are:

(1) You are hereby authorized to pay . . . .
(2) Passengers are warned to cross the track by the bridge only.

Indeed the verb may even be 'impersonal' in such cases with the passive, for example:

(3) Notice is hereby given that trespassers will be prosecuted.

This type is usually found on formal or legal occasions; and it is characteristic of it that, in writing at least, the word 'hereby' is often and perhaps can always be inserted; this serves to indicate that the utterance (in writing) of the sentence is, as it is said, the instrument effecting the act of warning, authorizing, &c. 'Hereby' is a useful criterion that the utterance is performative. If it is not

put in, 'passengers are . . .' may be used for the description of what usually happens; as for example in 'on nearing the tunnel, passengers are warned to duck their heads, &c.'

However, if we turn away from these highly formalized and explicit performative utterances, we have to recognize that mood and tense (hitherto retained as opposed to person and voice) break down as absolute criteria.

Mood (whatever this may be in English as opposed to Latin) will not do, for I may order you to turn right by saying, not 'I order you to turn right', but simply 'Turn right'; I may give you permission to go by saying simply 'You may go'; and instead of 'I advise [or "recommend"] you to turn right' I may say 'I should turn to the right if I were you'. Tense will not do either, for in giving (or calling) you off-side I may say, instead of 'I give [or "call"] you off-side', simply 'You were off-side'; and similarly, instead of saying 'I find you guilty' I may just say 'You did it'. Not to mention cases where we have only a truncated sentence, as when I accept a bet by saying simply 'Done', and even cases where there is no explicit verb at all, as when I say simply 'Guilty' in finding a person guilty, or 'Out' to give someone out.

Particularly with some special performative-looking words, for example 'off-side', 'liable', &c., we seem able to refute even the rule governing the use of the active or passive which we gave above. Instead of 'I pronounce you off-side' I might say 'You are off-side' and I might say 'I am (hereby rendered) liable' instead of 'I undertake . . .'. So we might think certain *words* might do

as a test of the performative utterance, that we could do it by means of *vocabulary* as distinct from *grammar*. Such words might be 'off-side', 'authorized', 'promise', 'dangerous', &c. But this will not do, for:

I. We may get the performative without the operative words thus:

(1) In place of 'dangerous corner' we may have 'corner', and in place of 'dangerous bull' we may write 'bull'.

(2) In place of 'you are ordered to . . .', we may have 'you will', and in place of 'I promise to . . .' we may have 'I shall'.

II. We may get the operative word without the utterance being performative, thus:

(1) In cricket a spectator may say 'it was over (really)'. Similarly I may say 'you were guilty' or 'you were off-side' or even 'you are guilty (off-side)' when I have no right to pronounce you guilty or off-side.

(2) In such locutions as 'you promised', 'you authorize' &c., the word occurs in a non-performative use.

This reduces us to an impasse over any *single simple* criterion of grammar or vocabulary. But maybe it is not impossible to produce a complex criterion, or at least a set of criteria, simple or complex, involving both grammar and vocabulary. For example, one of the criteria might be that everything with the verb in the imperative mood is performative (this leads, however, to

many troubles over, for example, when a verb is in the imperative mood and when it is not, into which I do not propose to go).

I would rather go back a minute and consider whether there was not some good reason behind our initial favouritism for verbs in the so-called 'present indicative active'.

We said that the idea of a performative utterance was that it was to be (or to be included as a part of) the performance of an action. Actions can only be performed by persons, and obviously in our cases the utterer must be the performer: hence our justifiable feeling—which we wrongly cast into purely grammatical mould—in favour of the 'first person', who must come in, being mentioned or referred to; moreover, if in uttering one is acting, one must be doing something—hence our perhaps ill-expressed favouring of the grammatical present and grammatical active of the verb. There is something which is *at the moment of uttering being done by the person uttering*.

Where there is *not*, in the verbal formula of the utterance, a reference to the person doing the uttering, and so the acting, by means of the pronoun 'I' (or by his personal name), then in fact he will be 'referred to' in one of two ways:

(*a*) In verbal utterances, *by his being the person who does* the uttering—what we may call the utterance-*origin* which is used generally in any system of verbal reference-co-ordinates.

(*b*) In written utterances (or 'inscriptions), *by his appending his signature* (this has to be done because, of

course, written utterances are not tethered to their origin in the way spoken ones are).

The 'I' who is doing the action does thus come essentially into the picture. An advantage of the original first person singular present indicative active form—or likewise of the second and third and impersonal passive forms with signature appended—is that this implicit feature of the speech-situation is made *explicit*. Moreover, the verbs which seem, on grounds of vocabulary, to be specially performative verbs serve the special purpose of *making explicit* (which is not the same as stating or describing) what precise action it is that is being performed by the issuing of the utterance: other words which seem to have a special performative function (and indeed *have* it), such as 'guilty', 'off-side', &c., do so because, in so far as and when they are linked in 'origin' with these special explicit performative verbs like 'promise', pronounce', 'find', &c.

The 'hereby' formula is a useful alternative; but it is rather too formal for ordinary purposes, and further, we may say 'I hereby state . . .' or 'I hereby question . . .', whereas we were hoping to find a criterion to distinguish statements from performatives. (I must explain again that we are floundering here. To feel the firm ground of prejudice slipping away is exhilarating, but brings its revenges.)

Thus what we should feel tempted to say is that any utterance which is in fact a performative should be reducible, or expandible, or analysable into a form, or

reproducible in a form, with a verb in the first person singular present indicative active (grammatical). This is the sort of test we were in fact using above. Thus:

'Out' is equivalent to 'I declare, pronounce, give, or call you out' (when it is a performative: it need not be, for example, if you are called out by someone not the umpire or recorded as 'out' by the scorer).

'Guilty' is equivalent to 'I find, pronounce, deem you to be guilty.'

'You are warned that the bull is dangerous' is equivalent to 'I, John Jones, warn you that the bull is dangerous' or

This bull is dangerous.
(Signed) John Jones.

This sort of expansion makes explicit both that the utterance is performative, and which act it is that is being performed. Unless the performative utterance is reduced to such an explicit form, it will regularly be possible to take it in a non-performative way: for example, 'it is yours' may be taken as equivalent to either 'I give it you' or 'it (already) belongs to you'. In fact there is rather a play on the performative and non-performative uses in the road sign 'You have been warned'.

However, though we might make progress along these lines (there are snags)[1] we must notice that this first

[1] For example, which are the verbs with which we can do this? If the performative is expanded, what is the test whether the first person singular present indicative active is on this occasion performative granted that all others have to be reducible (save the mark!) to this normal form?

person singular present indicative active, so called, is a *peculiar and special use*. In particular we must notice that there is an *asymmetry* of a systematic kind between it and other persons and tenses of the *very same verb*. The fact that there is *this* asymmetry is precisely the mark of the performative verb (and the nearest thing to a *grammatical* criterion in connexion with performatives).

Let us take an example: the uses of 'I bet' as opposed to the use of that verb in another tense or in another person. 'I betted' and 'he bets' are not performatives but describe actions on my and his part respectively—actions each consisting in the utterance of the performative 'I bet'. If I utter the words 'I bet . . .', I do not state that I utter the words 'I bet', or any other words, but I perform the act of betting; and similarly, if he says he bets, i.e. says the words 'I bet', he *bets*. But if I utter the words 'he bets', I only state that he utters (or rather has uttered) the words 'I bet': I do not perform his act of betting, which only he can perform: I describe his performances of the act of betting, but I do my own betting, and he must do his own. Similarly an anxious parent when his child has been asked to do something may say 'he promises, don't you Willy?' but little Willy must still himself say 'I promise' if he is really to have promised. Now this sort of asymmetry does not arise at all in general with verbs that are not used as explicit performatives. For example, there is no such asymmetry between 'I run' and 'He runs'.

Still, it is doubtful whether this is a 'grammatical'

criterion exactly (what is?), and anyway it is not very exact because:

(1) The first person singular present indicative active may be used to describe how I habitually behave: 'I bet him (every morning) sixpence that it will rain' or 'I promise only when I intend to keep my word'.

(2) The first person singular present indicative active may be used in a way similar to the 'historic' present. It may be used to describe my own performances elsewhere and elsewhen: 'on page 49 I protest against the verdict'. We might back this up by saying that performative verbs are not used in the present continuous tense (in the first person singular active): we do not say 'I am promising', and 'I am protesting'. But even this is not entirely true, because I can say 'Don't bother me at the moment; I will see you later; I am marrying' at any moment during the ceremony when I am not having to say other words such as 'I do'; here the utterance of the performative is not the whole of the performance, which is protracted and contains diverse elements. Or I can say 'I am protesting' when performing the act by, in this case, means other than saying 'I protest', for example by chaining myself to park railings. Or I can even say 'I am ordering' while writing the words 'I order'.

(3) Some verbs may be used in the first person singular present indicative active simultaneously in two ways. An example is 'I call', as when I say 'I call inflation too much money chasing too few goods' which embraces both a

performative utterance and a description of a naturally consequent performance.

(4) We shall be in apparent danger of bringing in many formulas which we might not like to class as performatives; for example 'I state that' (to utter which *is* to state) as well as 'I bet that'. In both examples there is the same asymmetry between first person and other uses.

(5) We have cases of suiting the action to the word: thus I may say 'I spit me of you' or *j'adoube* said when I give check, or 'I quote' followed by actually quoting. If I define by saying 'I define x as follows: x is y', this is a case of suiting the action (here giving a definition) to the word; when we use the formula 'I define x as y' we have a transition to a performative utterance from suiting the action to the word. We might add, too, that there is likewise a transition from the use of words as what we may call markers, to performatives. There is a transition from the word END at the end of a novel to the expression 'message ends' at the end of a signal message, to the expression 'with that I conclude my case' as said by Counsel in a law court. These, we may say, are cases of *marking* the action by the word, where eventually the use of the word comes to be the action of 'ending' (a difficult act to perform, being the cessation of acting, or to make explicit in other ways, of course).

(6) Is it always the case that we must have a performative verb for making explicit something we are undoubtedly doing by saying something? For example, I

may insult you by saying something, but we have not the formula 'I insult you'.

(7) Is it really the case that we can always put a performative into the normal form without loss? 'I shall . . .' can be meant in different ways; perhaps we trade on this. Or again we say 'I am sorry'; is this really exactly like the explicit 'I apologize'?

We shall have to revert to the notion of the explicit performative, and we must discuss *historically* at least how some of these perhaps not ultimately serious perplexities arise.

LECTURE VI

BECAUSE we suggested that the performative is not altogether so obviously distinct from the constative—the former happy or unhappy, the latter true or false—we were considering how to define the performative more clearly. The first suggestion was a criterion or criteria of grammar or of vocabulary or of both. We pointed out that there was certainly no one absolute criterion of this kind: and that very probably it is not possible to lay down even a list of all possible criteria; moreover, they certainly would not distinguish performatives from constatives, as very commonly the *same* sentence is used on different occasions of utterance in *both* ways, performative and constative. The thing seems hopeless from the start, if we are to leave utterances *as they stand* and seek for a criterion.

But nevertheless the type of performative upon which we drew for our first examples, which has a verb in the first person singular present indicative active, seems to deserve our favour: at least, if issuing the utterance is doing something, the 'I' and the 'active' and the 'present' seem appropriate. Though indeed performatives are not really like the remainder of the verbs in this 'tense' at all; there is an essential *asymmetry* with these verbs. This asymmetry is just the characteristic of a long list of

performative-looking verbs. The suggestion is, then, that we might

(1) make a list of all verbs with this peculiarity;
(2) suppose that all performative utterances which are not in fact in this preferred form—beginning 'I *x* that', 'I *x* to', or 'I *x*'—could be 'reduced' to this form and so rendered what we may call *explicit* performatives.

We are now asking: just how easy—even possible—is this going to be? It is fairly easy to make allowances for certain normal enough but different uses of the first person of the present indicative active even with these verbs, which may well be constative or descriptive, that is, the habitual present, the 'historic' (quasi-) present, and the continuous present. But then, as I was hastily mentioning, in conclusion, there are still further difficulties: we mentioned three as typical.

(1) 'I class' or perhaps 'I hold' seems in a way one, in a way the other. Which is it, or is it both?

(2) 'I state that' seems to conform to our grammatical or quasi-grammatical requirements: but do we want *it* in? Our criterion, such as it is, seems in danger of letting in non-performatives.

(3) Sometimes saying something seems to be characteristically doing something—for example insulting somebody, like reprimanding somebody: yet there is no performative 'I insult you'. Our criterion will not get in all cases of the issuing of an utterance being the

doing of something, because the 'reduction' to an explicit performative does not seem always possible.

Let us pause then to dwell a little more on the expression 'explicit performative', which we have introduced rather surreptitiously. I shall oppose it to 'primary performative' (rather than to inexplicit or implicit performative). We gave as an example:

(1) primary utterance: 'I shall be there',

(2) explicit performative: 'I promise that I shall be there', and we said that the latter formula made explicit what action it is that is being performed in issuing the utterance: i.e. 'I shall be there'. If someone says 'I shall be there', we might ask: 'Is that a promise?' We may receive the answer 'Yes', or 'Yes, I promise it' (or 'that . . .' or 'to . . .'), whereas the answer might have been only: 'No, but I do intend to be' (expressing or announcing an intention), or 'No, but I can foresee that, knowing my weaknesses, I (probably) shall be there'.

Now we must enter two caveats: 'making explicit' is not the same as describing or stating (at least in philosophers' preferred senses of these words) what I am doing. If 'making explicit' conveys this, then *pro tanto* it is a bad term. The situation in the case of actions which are non-linguistic but similar to performative utterances in that they are the performance of a conventional action (here ritual or ceremonial) is rather like this: suppose I bow deeply before you; it might not be clear whether I am doing obeisance to you or, say, stooping to observe the flora or to ease my indigestion. Generally speaking,

then, to make clear both *that* it is a conventional ceremonial act, and *which* act it is, the act (for example of doing obeisance) will as a rule include some special further feature, for example raising my hat, tapping my head on the ground, sweeping my other hand to my heart, or even very likely uttering some noise or word, for example 'Salaam'. Now uttering 'Salaam' is no more describing my performance, stating that I am performing an act of obeisance, than is taking off my hat: and by the same token (though we shall come back to this) saying 'I salute you' is no more describing my performance than is saying 'Salaam'. To do or to say these things is to make plain how the action is to be taken or understood, what action it is. And so it is with putting in the expression 'I promise that'. It is not a description, because (1) it could not be false, nor, therefore, true; (2) saying 'I promise that' (if happy, of course) *makes it* a promise, and *makes it* unambiguously a promise. Now we can say that such a performative formula as 'I promise' makes it clear how what is said is to be understood and even conceivably that the formula 'states that' a promise has been made; but we cannot say that such utterances are true or false, nor that they are descriptions or reports.

Secondly, a minor caution: notice that, although we have in this type of utterance a 'that-' clause following a verb, for example 'promise', or 'find', or 'pronounce' (or perhaps such verbs as 'estimate'), we must not allude to this as 'indirect speech'. 'That'-clauses in indirect speech or *oratio obliqua* are of course cases where I report what

someone else or myself elsewhen or elsewhere did say: for example, typically, 'he said that . . .', but also possibly 'he promised that . . .' (or is this a double use of 'that'?), or 'on page 456 I declared that . . .'. If this is a clear notion[1] we see that the 'that' of *oratio obliqua* is not in all ways similar to the 'that' in our explicit performative formulas: here I am not reporting my own speech in the first person singular present indicative active. Incidentally, of course, it is not in the least necessary that an explicit performative verb should be followed by 'that': in important classes of cases it is followed by 'to . . .' or nothing, for example, 'I apologize (for . . .)', 'I salute you'.

Now, one thing that seems at least a fair guess, even from the elaboration of the linguistic construction, as also from its nature in the explicit performative is this: that historically, from the point of view of the evolution of language, the explicit performative must be a later development than certain more primary utterances, many of which at least are already implicit performatives, which are included in most or many explicit performatives as parts of a whole. For example, 'I will . . .' is earlier than 'I promise that I will . . .'. The plausible view (I do not know exactly how it would be established) would be that in primitive languages it would not yet be clear, it would not yet be possible to distinguish, which of various things that (using later distinctions) we might be doing

[1] My explanation is very obscure, like those of all grammar books on 'that' clauses: compare their even worse explanation of 'what' clauses.

we were in fact doing. For example 'Bull' or 'Thunder' in a primitive language of one-word utterances[1] could be a warning, information, a prediction, &c. It is also a plausible view that explicitly distinguishing the different *forces* that this utterance might have is a later achievement of language, and a considerable one; primitive or primary forms of utterance will preserve the 'ambiguity' or 'equivocation' or 'vagueness' of primitive language in this respect; they will not make explicit the precise force of the utterance. This may have its uses: but sophistication and development of social forms and procedures will necessitate clarification. But note that this clarification is as much a creative act as a discovery or description! It is as much a matter of making clear distinctions as of making already existent distinctions clear.

One thing, however, that it will be most dangerous to do, and that we are very prone to do, is to take it that we somehow *know* that the primary or primitive use of sentences must be, because it ought to be, statemental or constative, in the philosophers' preferred sense of simply uttering something whose sole pretension is to be true or false and which is not liable to criticism in any other dimension. We certainly do not know that this is so, any more, for example, than that all utterances must have first begun as imperatives (as some argue) or as swear-words—and it seems much more likely that the 'pure' statement is a goal, an ideal, towards which the gradual development of science has given the impetus, as it has

[1] As in fact primitive languages probably were, cf. Jespersen.

likewise also towards the goal of precision. Language as such and in its primitive stages is not precise, and it is also not, in our sense, explicit: precision in language makes it clearer what is being said—its *meaning*: explicitness, in our sense, makes clearer the *force* of the utterances, or 'how (in one sense; see below) it is to be taken'.

The explicit performative formula, moreover, is only the last and 'most successful' of numerous speech-devices which have always been used with greater or less success to perform the same function (just as measurement or standardization was the most successful device ever invented for developing *precision* of speech).

Consider for a moment *some* of these other more primitive devices in speech, some of the roles which can (though, of course, not without change or loss, as we shall see) be taken over by the device of the explicit performative.

1. *Mood*

We have already mentioned the exceedingly common device of using the imperative mood. This makes the utterance a 'command' (or an exhortation or permission or concession or what not!) Thus I may say 'shut it' in many contexts:

'Shut it, do' resembles 'I order you to shut it'.

'Shut it—I should' resembles 'I advise you to shut it'.

'Shut it, if you like' resembles 'I permit you to shut it'.

'Very well then, shut it' resembles 'I consent to your shutting it'.

'Shut it if you dare' resembles 'I dare you to shut it'.

Or again we may use auxiliaries:

'You may shut it' resembles 'I give permission, I consent, to your shutting it'.

'You must shut it' resembles 'I order you, I advise you, to shut it'.

'You ought to shut it' resembles 'I advise you to shut it'.

2. *Tone of voice, cadence, emphasis*

(Similar to this is the sophisticated device of using stage directions; for example, 'threateningly', &c.) Examples of this are:

It's going to charge! (a warning);
It's going to charge? (a question);
It's going to charge!? (a protest).

These features of spoken language are not reproducible readily in written language. For example we have tried to convey the tone of voice, cadence and emphasis of a protest by the use of an exclamation mark and a question mark (but this is very jejune). Punctuation, italics, and word order may help, but they are rather crude.

3. *Adverbs and adverbial phrases*

But in written language—and even, to some extent, in spoken language, though there they are not so necessary—we rely on adverbs, adverbial phrases, or turns of

phrase. Thus we can qualify the force of 'I shall' by adding 'probably' or—in an opposite sense—by adding 'without fail'; we can give emphasis (to a reminder or whatever it may be) by writing 'You would do well never to forget that . . .'. Much could be said about the connexions here with the phenomena of evincing, intimating, insinuation, innuendo, giving to understand, enabling to infer, conveying, 'expressing' (odious word) all of which are, however, essentially different, though they involve the employment of very often the same or similar verbal devices and circumlocutions. In the latter half of our lectures we shall revert to the important and difficult distinction which needs to be drawn here.

4. *Connecting particles*

At a more sophisticated level, perhaps, comes the use of the special verbal device of the connecting particle; thus we may use the particle 'still' with the force of 'I insist that'; we use 'therefore' with the force of 'I conclude that'; we use 'although' with the force of 'I concede that'. Note also the uses of 'whereas' and 'hereby' and 'moreover'.[1] A very similar purpose is served by the use of titles such as Manifesto, Act, Proclamation, or the subheading 'A Novel . . .'.

Moreover, even apart from and turning from what we say and the manner of speaking it, there are other

[1] But some of these examples raise the old question whether 'I concede that' and 'I conclude that' are performatives or not.

essential devices by which the force of the utterance is to some extent got across:

5. *Accompaniments of the utterance*

We may accompany the utterance of the words by gestures (winks, pointings, shruggings, frowns, &c.) or by ceremonial non-verbal actions. These may sometimes serve without the utterance of any words, and their importance is very obvious.

6. *The circumstances of the utterance*

An exceedingly important aid is the circumstances of the utterance. Thus we may say 'coming from *him*, I took it as an order, not as a request'; similarly the context of the words 'I shall die some day', 'I shall leave you my watch', in particular the health of the speaker, make a difference how we shall understand them.

But in a way these resources are over-rich: they lend themselves to equivocation and inadequate discrimination; and moreover, we use them for other purposes, e.g. insinuation. The explicit performative rules out equivocation and keeps the performance fixed, relatively.

The trouble about all these devices has been principally their vagueness of meaning and uncertainty of sure reception, but there is also probably some positive inadequacy in them for dealing with anything like the complexity of the field of actions which we perform with words. An 'imperative' may be an order, a permission, a demand, a request, an entreaty, a suggestion, a recommendation, a warning ('go and you will see'), or may express a condition or concession or a definition ('Let it . . .'), &c. To hand something over to someone may be, when we say 'Take it', the giving it or lending it or leasing it or entrusting it. To say 'I shall' may be to promise, or to express an intention, or to forecast my future. And so on. No doubt a combination of some or all the devices mentioned above (and very likely there are others) will usually, if not in the end, suffice. Thus when we say 'I shall' we can make it clear that we are forecasting by adding the adverbs 'undoubtedly' or 'probably', that we are expressing an intention by adding the adverbs 'certainly' or 'definitely', or that we are promising by adding the adverbial phrase 'without fail', or saying 'I shall do my best to'.

It should be noted that when performative verbs exist we can use them not only in 'that . . .' or 'to . . .' formulas, but also in stage directions ('welcomes'), titles ('warning!'), and parentheses (this is almost as good a test of a performative as our normal forms); and we must not forget the use of special words such as 'Out', &c., which have no normal form.

However, the existence and even the use of explicit performatives does not remove all our troubles.

(1) In philosophy, we can even raise the trouble of the liability of performatives to be mistaken for descriptives or constatives.

(1*a*) Nor, of course, is it merely that the performative does not preserve the often congenial equivocation of

primary utterances; we must also in passing consider cases where it is doubtful whether the expression is an explicit performative or not and cases very similar to performatives but not performatives.

(2) There seem to be clear cases where the very same formula seems sometimes to be an explicit performative and sometimes to be a descriptive, and may even trade on this ambivalence: for example, 'I approve' and 'I agree'. Thus 'I approve' may have the performative force of giving approval or it may have a descriptive meaning: 'I favour this'.

We shall consider two classic sorts of case in which this will arise. They exhibit some of the phenomena incidental to the development of explicit performative formulas.

There are numerous cases in human life where the feeling of a certain 'emotion' (save the word!) or 'wish' or the adoption of an attitude is conventionally considered an appropriate or fitting response or reaction to a certain state of affairs, including the performance by someone of a certain act, cases where such a response is natural (or we should like to think so!) In such cases it is, of course, possible and usual actually to feel the emotion or wish in question; and since our emotions or wishes are not readily detectable by others, it is common to wish to inform others that we have them. Understandably, though for slightly different and perhaps less estimable reasons in different cases, it becomes *de rigueur* to 'express' these feelings if we have them, and further even to

express them when they are felt fitting, regardless of whether we really feel anything at all which we are reporting. Examples of expressions so used are:

I thank	I am grateful	I feel grateful
I apologize	I am sorry	I repent
I criticize / I censure	I blame	I am shocked by / I am revolted by
I approve	I approve of	I feel approval
I bid you welcome	I welcome	
I congratulate	I am glad about	

In these lists, the first column contains performative utterances; those in the second are not pure but half descriptive, and in the third are merely reports. There are then here numerous expressions, among them many important ones, which suffer from or profit by a sort of deliberate ambivalence, and this is fought by the constant introduction of deliberately pure performative phrases. Can we suggest any tests for deciding whether 'I approve of' or 'I am sorry' is being used (or even is always used) in the one way or the other?

One test would be whether it makes sense to say 'Does he *really*?' For example, when someone says 'I welcome you' or 'I bid you welcome', we may say 'I wonder if he really did welcome him?' though we could not say in the same way 'I wonder whether he really does bid him welcome?' Another test would be whether one could really be doing it without actually saying anything, for example in the case of being sorry as distinct from apologizing, in

being grateful as distinct from thanking, in blaming as distinct from censuring.[1] Yet a third test would be, at least in some cases, to ask whether we could insert before the supposed performative verb some such adverb as 'deliberately' or such an expression as 'I am willing to': because (possibly) if the utterance is the doing of an action, then it is surely something we ought to be able (on occasion) to do deliberately or to be willing to do. Thus we may say: 'I deliberately bade him welcome', 'I deliberately approved his action', 'I deliberately apologized', and we can say 'I am willing to apologize'. But we cannot say 'I deliberately approved of his action' or 'I am willing to be sorry' (as distinct from 'I am willing to say I am sorry').

A fourth test would be to ask whether what one says could be literally false, as sometimes when I say 'I am sorry', or could only involve insincerity (unhappiness) as sometimes when I say 'I apologize': these phrases blur the distinction between insincerity and falsehood.[2]

But there is here a certain distinction to be drawn in passing of the exact nature of which I am uncertain: we have related 'I apologize' to 'I am sorry' as above; but now there are also very numerous conventional expressions of feeling, very similar in some ways, which are

[1] There are classic doubts about the possibility of tacit consent; here non-verbal performance occurs in an alternative form of performative act: this casts doubt on this second test!

[2] There are parallel phenomena to these in other cases: for example a specially confusing one arises over what we may call dictional or expositive performatives.

certainly nothing to do with performatives: for example:

'I have pleasure in calling upon the next speaker'.
'I am sorry to have to say . . .'.
'I am gratified to be in a position to announce . . .'.[1]

We may call these *polite* phrases, like 'I have the honour to . . .'. It is conventional enough to formulate them in this way: but it is *not* the case that to say you have pleasure in *is* to have pleasure in doing something. Unfortunately. To be a performative utterance, even in these cases connected with feelings and attitudes which I christen 'BEHABITIVES', is not *merely* to be a conventional expression of feeling or attitude.

Also to be distinguished are cases of *suiting the action to the word*—a special type of case which may generate performatives but which is not in itself a case of the performative utterance. A typical case is: 'I slam the door thus' (he slams the door). But this sort of case leads to 'I salute you' (he salutes); here 'I salute you' may become a substitute for the salute and thus a pure performative utterance. To say 'I salute you' now *is* to salute you. Compare the expression 'I salute the memory . . .'.

But there are many transitional stages between suiting the action to the word and the pure performative:

'Snap.' To say this is to snap (in appropriate circumstances); but it is not a snap if 'snap' is not said.

[1] [Marginal note in manuscript: 'Further classification needed here: just note it in passing.']

'Check.' To say it is to check in appropriate circumstances. But would it not still be a check if 'check' were not said?

'J'adoube.' Is this suiting the action to the word or is it part of the act of straightening the piece as opposed to moving it?

Perhaps these distinctions are not important: but there are similar transitions in the case of performatives, as for example:

'I quote': he quotes.
'I define': he defines (e.g. *x* is *y*).
'I define *x* as *y*'.

In these cases the utterance operates like a title: is it a variety of performative? It essentially operates where the action suited to the word is itself a verbal performance.

LECTURE VIII

IN embarking on a programme of finding a list of explicit performative verbs, it seemed that we were going to find it not always easy to distinguish performative utterances from constative, and it therefore seemed expedient to go farther back for a while to fundamentals—to consider from the ground up how many senses there are in which to say something *is* to do something, or *in* saying something we do something, and even *by* saying something we do something. And we began by distinguishing a whole group of senses of 'doing something' which are all included together when we say, what is obvious, that to say something is in the full normal sense to do something—which includes the utterance of certain noises, the utterance of certain words in a certain construction, and the utterance of them with a certain 'meaning' in the favourite philosophical sense of that word, i.e. with a certain sense and with a certain reference.

The act of 'saying something' in this full normal sense I call, i.e. dub, the performance of a locutionary act, and the study of utterances thus far and in these respects the study of locutions, or of the full units of speech. Our interest in the locutionary act is, of course, principally to make quite plain what it is, in order to distinguish it from other acts with which we are going to be primarily

concerned. Let me add merely that, of course, a great many further refinements would be possible and necessary if we were to discuss it for its own sake—refinements of very great importance not merely to philosophers but to, say, grammarians and phoneticians.

We had made three rough distinctions between the phonetic act, the phatic act, and the rhetic act. The phonetic act is merely the act of uttering certain noises. The phatic act is the uttering of certain vocables or words, i.e. noises of certain types, belonging to and as belonging to, a certain vocabulary, conforming to and as conforming to a certain grammar. The rhetic act is the performance of an act of using those vocables with a certain more-or-less definite sense and reference. Thus 'He said "The cat is on the mat" ', reports a phatic act, whereas 'He said that the cat was on the mat' reports a rhetic act. A similar contrast is illustrated by the pairs:

'He said "The cat is on the mat" ', 'He said (that) the cat was on the mat';
'He said "I shall be there" ', 'He said he would be there';
'He said "Get out" ', 'He told me to get out';
'He said "Is it in Oxford or Cambridge?" '; 'He asked whether it was in Oxford or Cambridge'.

To pursue this for its own sake beyond our immediate requirements, I shall mention some general points worth remembering:

(1) Obviously, to perform a phatic I must perform a

phonetic act, or, if you like, in performing one I am performing the other (not, however, that phatic acts are a sub-class of phonetic acts; we defined the phatic act as the uttering of vocables *as* belonging to a certain vocabulary): but the converse is not true, for if a monkey makes a noise indistinguishable from 'go' it is still not a phatic act.

(2) Obviously in the definition of the phatic act two things were lumped together: vocabulary and grammar. So we have not assigned a special name to the person who utters, for example, 'cat thoroughly the if' or 'the slithy toves did gyre'. Yet a further point arising is the intonation as well as grammar and vocabulary.

(3) The phatic act, however, like the phonetic, is essentially mimicable, reproducible (including intonation, winks, gestures, &c.). One can mimic not merely the statement in quotation marks 'She has lovely hair', but also the more complex fact that he said it like this: 'She has lovely *hair*' (shrugs).

This is the 'inverted commas' use of 'said' as we get it in novels: every utterance can be just reproduced in inverted commas, or in inverted commas with 'said he' or, more often, 'said she', &c., after it.

But the rhetic act is the one we report, in the case of assertions, by saying 'He said that the cat was on the mat', 'He said he would go', 'He said I was to go' (his words were 'You are to go'). This is the so-called 'indirect speech'. If the sense or reference is *not* being taken as clear, then the whole or part is to be in quotation marks. Thus

I might say: 'He said I was to go to "the minister", but he did not say which minister' or 'I said that he was behaving badly and he replied that "the higher you get the fewer" '. We cannot, however, always use 'said that' easily: we would say 'told to', 'advise to', &c., if he used the imperative mood, or such equivalent phrases as 'said I was to', 'said I should', &c. Compare such phrases as 'bade me welcome' and 'extended his apologies'.

I add one further point about the rhetic act: of course sense and reference (naming and referring) themselves are here ancillary acts performed in performing the rhetic act. Thus we may say 'I meant by "bank" . . .' and we say 'by "he" I was referring to . . .'. Can we perform a rhetic act without referring or without naming? In general it would seem that the answer is that we cannot, but there are puzzling cases. What is the reference in 'all triangles have three sides'? Correspondingly, it is clear that we can perform a phatic act which is not a rhetic act, though not conversely. Thus we may repeat someone else's remark or mumble over some sentence, or we may read a Latin sentence without knowing the meaning of the words.

The question when one pheme or one rheme is the *same* as another, whether in the 'type' or 'token' sense, and the question what is one single pheme or rheme, do not so much matter here. But, of course, it is important to remember that the same pheme, e.g., sentence, that is, tokens of the same type, may be used on different occasions of utterance with a different sense or reference,

and so be a different rheme. When different phemes are used with the same sense and reference, we might speak of rhetically equivalent acts ('the same statement' in one sense) but not of the same rheme or rhetic acts (which are the same statement in another sense which involves using the same words).

The pheme is a unit of *language*: its typical fault is to be nonsense—meaningless. But the rheme is a unit of *speech*; its typical fault is to be vague or void or obscure, &c.

But though these matters are of much interest, they do not so far throw any light at all on our problem of the constative as opposed to the performative utterance. For example, it might be perfectly possible, with regard to an utterance, say 'It is going to charge', to make entirely plain 'what we were saying' in issuing the utterance, in all the senses so far distinguished, and yet not at all to have cleared up whether or not in issuing the utterance I was performing the act of *warning* or not. It may be perfectly clear what I mean by 'It is going to charge' or 'Shut the door', but not clear whether it is meant as a statement or warning, &c.

To perform a locutionary act is in general, we may say, also and *eo ipso* to perform an *illocutionary* act, as I propose to call it. Thus in performing a locutionary act we shall also be performing such an act as:

asking or answering a question,
giving some information or an assurance or a warning,
announcing a verdict or an intention,

pronouncing sentence,
making an appointment or an appeal or a criticism,
making an identification or giving a description,

and the numerous like. (I am not suggesting that this is a clearly defined class by any means.) There is nothing mysterious about our *eo ipso* here. The trouble rather is the number of different senses of so vague an expression as 'in what way are we using it'—this may refer even to a locutionary act, and further to perlocutionary acts to which we shall come in a minute. When we perform a locutionary act, we use speech: but in what way precisely are we using it on this occasion? For there are very numerous functions of or ways in which we use speech, and it makes a great difference to our act in some sense—sense (B)[1]—in which way and which *sense* we were on this occasion 'using' it. It makes a great difference whether we were advising, or merely suggesting, or actually ordering, whether we were strictly promising or only announcing a vague intention, and so forth. These issues penetrate a little but not without confusion into grammar (see above), but we constantly do debate them, in such terms as whether certain words (a certain locution) *had the force of* a question, or *ought to have been taken as* an estimate and so on.

I explained the performance of an act in this new and second sense as the performance of an 'illocutionary' act, i.e. performance of an act *in* saying something as opposed

[1] See below, p. 101.

to performance of an act *of* saying something; I call the act performed an 'illocution' and shall refer to the doctrine of the different types of function of language here in question as the doctrine of 'illocutionary forces'.

It may be said that for too long philosophers have neglected this study, treating all problems as problems of 'locutionary usage', and indeed that the 'descriptive fallacy' mentioned in Lecture I commonly arises through mistaking a problem of the former kind for a problem of the latter kind. True, we are now getting out of this; for some years we have been realizing more and more clearly that the occasion of an utterance matters seriously, and that the words used are to some extent to be 'explained' by the 'context' in which they are designed to be or have actually been spoken in a linguistic interchange. Yet still perhaps we are too prone to give these explanations in terms of 'the meanings of words'. Admittedly we can use 'meaning' also with reference to illocutionary force—'He meant it as an order', &c. But I want to distinguish *force* and meaning in the sense in which meaning is equivalent to sense and reference, just as it has become essential to distinguish sense and reference.

Moreover, we have here an illustration of the different uses of the expression, 'uses of language', or 'use of a sentence', &c.—'use' is a hopelessly ambiguous or wide word, just as is the word 'meaning', which it has become customary to deride. But 'use', its supplanter, is not in much better case. We may entirely clear up the 'use of a sentence' on a particular occasion, in the sense of the

locutionary act, without yet touching upon its use in the sense of an *illocutionary* act.

Before refining any further on this notion of the illocutionary act, let us contrast both the locutionary *and* the illocutionary act with yet a third kind of act.

There is yet a further sense (C) in which to perform a locutionary act, and therein an illocutionary act, may also be to perform an act of another kind. Saying something will often, or even normally, produce certain consequential effects upon the feelings, thoughts, or actions of the audience, or of the speaker, or of other persons: and it may be done with the design, intention, or purpose of producing them; and we may then say, thinking of this, that the speaker has performed an act in the nomenclature of which reference is made either (C. *a*), only obliquely, or even (C. *b*), not at all, to the performance of the locutionary or illocutionary act. We shall call the performance of an act of this kind the performance of a 'perlocutionary' act, and the act performed, where suitable—essentially in cases falling under (C. *a*)—a 'perlocution'. Let us not yet define this idea any more carefully—of course it needs it—but simply give examples:

(E. 1)

Act (A) or Locution

He said to me 'Shoot her!' meaning by 'shoot' shoot and referring by 'her' to *her*.

Act (B) or Illocution

He urged (or advised, ordered, &c.) me to shoot her.

Act (C. *a*) or Perlocution

He persuaded me to shoot her.

Act (C. *b*)

He got me to (or made me, &c.) shoot her.

(E. 2)

Act (A) or Locution

He said to me, 'You can't do that'.

Act (B) or Illocution

He protested against my doing it.

Act (C. *a*) or Perlocution

He pulled me up, checked me.

Act (C. *b*)

He stopped me, he brought me to my senses, &c.
He annoyed me.

We can similarly distinguish the locutionary act 'he said that . . .' from the illocutionary act 'he argued that . . .' and the perlocutionary act 'he convinced me that . . .'

It will be seen that the 'consequential effects' here mentioned (see C. *a* and C. *b*) do not include a particular kind of consequential effects, those achieved, e.g., by way

of committing the speaker as in promising, which come into the illocutionary act. Perhaps restrictions need making, as there is clearly a difference between what we feel to be the real production of real effects and what we regard as mere conventional consequences; we shall in any case return later to this.

We have here then roughly distinguished three kinds of acts—the locutionary, the illocutionary, and the perlocutionary. Let us make some general comments on these three classes, leaving them still fairly rough. The first three points will be about 'the use of language' again.

(1) Our interest in these lectures is essentially to fasten on the second, illocutionary act and contrast it with the other two. There is a constant tendency in philosophy to elide this in favour of one or other of the other two. Yet it is distinct from both. We have already seen how the expressions 'meaning' and 'use of sentence' can blur the distinction between locutionary and illocutionary acts. We now notice that to speak of the 'use' of language can likewise blur the distinction between the illocutionary and perlocutionary act—so we will distinguish them more carefully in a minute. Speaking of the 'use of "language" for arguing or warning' looks just like speaking of 'the use of "language" for persuading, rousing, alarming'; yet the former may, for rough contrast, be said to be *conventional*, in the sense that at least it could be made explicit by the performative formula; but the latter could not. Thus we can say 'I argue that' or 'I

warn you that' but we cannot say 'I convince you that' or 'I alarm you that'. Further, we may entirely clear up whether someone was arguing or not without touching on the question whether he was convincing anyone or not.

(2) To take this farther, let us be quite clear that the expression 'use of language' can cover other matters even more diverse than the illocutionary and perlocutionary acts and obviously quite diverse from any with which we are here concerned. For example, we may speak of the 'use of language' *for* something, e.g. for joking; and we may use 'in' in a way different from the illocutionary 'in', as when we say 'in saying "p" I was joking' or 'acting a part' or 'writing poetry'; or again we may speak of 'a poetical use of language' as distinct from 'the use of language in poetry'. These references to 'use of language' have nothing to do with the illocutionary act. For example, if I say 'Go and catch a falling star', it may be quite clear what both the meaning and the force of my utterance is, but still wholly unresolved which of these other kinds of things I may be doing. There are aetiolations, parasitic uses, etc., various 'not serious' and 'not full normal' uses. The normal conditions of reference may be suspended, or no attempt made at a standard perlocutionary act, no attempt to make you do anything, as Walt Whitman does not seriously incite the eagle of liberty to soar.

(3) Furthermore, there may be some things we 'do' in some connexion with saying something which do not seem to fall, intuitively at least, exactly into any of these roughly defined classes, or else seem to fall vaguely into

more than one; but any way we do not at the outset feel so clear that they are as remote from our three acts as would be joking or writing poetry. For example, *insinuating*, as when we insinuate something in or by issuing some utterance, seems to involve some convention, as in the illocutionary act; but we cannot *say* 'I insinuate . . .', and it seems like implying to be a clever effect rather than a mere act. A further example is evincing emotion. We may evince emotion in or by issuing an utterance, as when we swear; but once again we have no use here for performative formulas and the other devices of illocutionary acts. We might say that we use swearing[1] *for* relieving our feelings. We must notice that the illocutionary act is a conventional act: an act done as conforming to a convention.

The next three points that arise do so importantly because our acts are *acts*.

(4) Acts of all our three kinds necessitate, since they are the performing of actions, allowance being made for the ills that all action is heir to. We must systematically be prepared to distinguish between 'the act of doing *x*', i.e. achieving *x*, and 'the act of attempting to do *x*'.

In the case of illocutions we must be ready to draw the necessary distinction, not noticed by ordinary language except in exceptional cases, between

(*a*) the act of attempting or purporting (or affecting or professing or claiming or setting up or setting out) to perform a certain illocutionary act, and

[1] 'Swearing' is ambiguous: 'I swear by Our Lady' *is* to swear by Our Lady: but 'Bloody' is not to swear by Our Lady.

(*b*) the act of successfully achieving or consummating or bringing off such an act.

This distinction is, or should be, a commonplace of the theory of our language about 'action' in general. But attention has been drawn earlier to its special importance in connexion with performatives: it is always possible, for example, to try to thank or inform somebody yet in different ways to fail, because he doesn't listen, or takes it as ironical, or wasn't responsible for whatever it was, and so on. This distinction will arise, as over any act, over locutionary acts too; but failures here will not be unhappinesses as there, but rather failures to get the words out, to express ourselves clearly, etc.

(5) Since our acts are actions, we must always remember the distinction between producing effects or consequences which are intended or unintended; and (i) when the speaker intends to produce an effect it may nevertheless not occur, and (ii) when he does not intend to produce it or intends not to produce it it may nevertheless occur. To cope with complication (i) we invoke as before the distinction between attempt and achievement; to cope with complication (ii) we invoke the normal linguistic devices of disclaiming (adverbs like 'unintentionally' and so on) which we hold ready for general use in all cases of doing actions.[1]

[1] This complication (ii), it may be pointed out, can of course also arise in the cases of both locutionary and illocutionary acts. I may say something or refer to something without meaning to, or commit myself unintentionally to a certain undertaking; for example, I may order someone to do something, when I did not intend to order him to do so. But it is in connexion with perlocution that it is most prominent, as is also the distinction between attempt and achievement.

(6) Furthermore, we must, of course, allow that as actions they may be things that we do not exactly *do*, in the sense that we did them, say, under duress or in any other such way. Other ways besides in which we may not fully do the action are given in (2) above. We may, perhaps, add the cases given in (5) where we produce consequences by mistake, did not intend to do so.

(7) Finally we must meet the objection about our illocutionary and perlocutionary acts—namely that the notion of an act is unclear—by a general doctrine about action. We have the idea of an 'act' as a fixed physical thing that we do, as distinguished from conventions and as distinguished from consequences. But

(*a*) the illocutionary act and even the locutionary act too involve conventions: compare with them the act of doing obeisance. It is obeisance only because it is conventional and it is done only because it is conventional. Compare also the distinction between kicking a wall and kicking a goal;

(*b*) the perlocutionary act always includes some consequences, as when we say 'By doing *x* I was doing *y*': we do bring in a greater or less stretch of 'consequences' always, some of which may be 'unintentional'. There is no restriction to the minimum physical act at all. That we can import an arbitrarily long stretch of what might also be called the 'consequences' of our act into the nomenclature of the act itself is, or should be, a fundamental commonplace of the theory of our language about all 'action' in general. Thus if asked 'What did he do?', we may reply either 'He shot the donkey' or 'He fired a

gun' or 'He pulled the trigger' or 'He moved his trigger finger', and all may be correct. So, to shorten the nursery story of the endeavours of the old woman to drive her pig home in time to get her old man's supper, we may in the last resort say that the cat drove or got the pig, or made the pig get, over the stile. If in such cases we *mention* both a B act (illocution) and a C act (perlocution) we shall say '*by* B-ing he C-ed' rather than '*in*-B-ing . . .' This is the reason for calling C a *per*locutionary act as distinct from an illocutionary act.

Next time we shall revert to the distinction between our three kinds of act, and to the expressions 'in' and 'by doing *x* I am doing *y*', with a view to getting the three classes and their members and non-members somewhat clearer. We shall see that just as the locutionary act embraces doing many things at once to be complete, so may the illocutionary and perlocutionary acts.

Frege, G.

In: Mind 14 (1905)

ÜBER SINN UND BEDEUTUNG

Die Gleichheit[1] fordert das Nachdenken heraus durch Fragen, die sich daran knüpfen und nicht ganz leicht zu beantworten sind. Ist sie eine Beziehung? eine Beziehung zwischen Gegenständen? oder zwischen Namen oder Zeichen für Gegenstände? Das letzte hatte ich in meiner Begriffsschrift angenommen. Die Gründe, die dafür zu sprechen scheinen, sind folgende: $a = a$ und $a = b$ sind offenbar Sätze von verschiedenem Erkenntniswerte: $a = a$ gilt a priori und ist nach Kant analytisch zu nennen, während Sätze von der Form $a = b$ oft sehr wertvolle Erweiterungen unserer Erkenntnis enthalten und a priori nicht immer zu begründen sind. Die Entdeckung, daß nicht jeden Morgen eine neue Sonne aufgeht, sondern immer dieselbe, ist wohl eine der folgenreichsten in der Astronomie gewesen. Noch jetzt ist die Wiedererkennung eines kleinen Planeten oder eines Kometen nicht immer etwas Selbst- | verständliches. Wenn wir nun in der Gleichheit eine Beziehung zwischen dem sehen wollten, was die Namen »a« und »b« bedeuten, so schiene $a = b$ von $a = a$ nicht verschieden sein zu können, falls nämlich $a = b$ wahr ist. Es wäre hiermit eine Beziehung eines Dinges zu sich selbst ausgedrückt, und zwar eine solche, in der jedes Ding mit sich selbst, aber kein Ding mit einem andern steht. Was man mit $a = b$ sagen will, scheint zu sein, daß die Zeichen oder Namen »a« und »b« dasselbe bedeuten, und dann wäre eben von jenen Zeichen die Rede; es würde eine Beziehung zwischen ihnen behauptet. Aber diese Beziehung bestände zwischen den Namen oder Zeichen nur, insofern sie etwas benennen oder bezeichnen. Sie wäre eine vermittelte durch die Verknüpfung jedes der beiden Zeichen mit demselben Bezeichneten. Diese aber ist willkürlich. Man kann keinem verbieten, irgendeinen willkürlich hervorzubringenden Vorgang oder Gegenstand zum Zeichen für irgend etwas anzunehmen. Damit würde dann ein Satz $a = b$ nicht mehr die Sache selbst, sondern nur noch unsere Bezeichnungsweise betreffen; wir würden keine eigentliche Erkenntnis darin ausdrücken. Das wollen wir aber doch grade in vielen Fällen. Wenn sich das Zeichen »a« von dem Zeichen »b« nur als Gegenstand (hier durch die Gestalt) unterscheidet, nicht als Zeichen; das soll heißen: nicht in der Weise, wie es etwas bezeichnet: so würde der Erkenntniswert von $a = a$ wesentlich gleich dem von $a = b$ sein, falls $a = b$ wahr ist. Eine Verschiedenheit kann nur dadurch zustande kommen, daß der Unterschied des Zeichens einem Unterschiede in der Art des

Zeitschrift für Philosophie und philosophische Kritik. 100 (1892), S. 25–50.

[1] Ich brauche dies Wort im Sinne von Identität und verstehe »$a = b$« in dem Sinne von »a ist dasselbe wie b« oder »a und b fallen zusammen«.

Gegebenseins des Bezeichneten entspricht. Es seien a, b, c die Geraden, welche die Ecken eines Dreiecks mit den Mitten der Gegenseiten verbinden. Der Schnittpunkt von a und b ist dann derselbe wie der Schnittpunkt von b und c. Wir haben also verschiedene Bezeichnungen für denselben Punkt, und diese Namen (»Schnittpunkt von a und b« »Schnittpunkt von b und c«) deuten zugleich auf die Art des Gegebenseins, und daher ist in dem Satze eine wirkliche Erkenntnis enthalten.

Es liegt nun nahe, mit einem Zeichen (Namen, Wortverbindung, Schriftzeichen) außer dem Bezeichneten, was die Bedeutung des Zeichens heißen möge, noch das verbunden zu denken, was ich den Sinn des Zeichens nennen möchte, worin die Art des Gegebenseins enthalten ist. Es würde danach in unserm Beispiele zwar die | Bedeutung der Ausdrücke »der Schnittpunkt von a und b« und »der Schnittpunkt von b und c« dieselbe sein, aber nicht ihr Sinn. Es würde die Bedeutung von »Abendstern« und »Morgenstern« dieselbe sein, aber nicht der Sinn.

Aus dem Zusammenhange geht hervor, daß ich hier unter »Zeichen« und »Namen« irgendeine Bezeichnung verstanden habe, die einen Eigennamen vertritt, deren Bedeutung also ein bestimmter Gegenstand ist (dies Wort im weitesten Umfange genommen), aber kein Begriff und keine Beziehung, auf die in einem anderen Aufsatze näher eingegangen werden soll. Die Bezeichnung eines einzelnen Gegenstandes kann auch aus mehreren Worten oder sonstigen Zeichen bestehn. Der Kürze wegen mag jede solche Bezeichnung Eigenname genannt werden.

Der Sinn eines Eigennamens wird von jedem erfaßt, der die Sprache oder das Ganze von Bezeichnungen hinreichend kennt, der er angehört[2]; damit ist die Bedeutung aber, falls sie vorhanden ist, doch immer nur einseitig beleuchtet. Zu einer allseitigen Erkenntnis der Bedeutung würde gehören, daß wir von jedem gegebenen Sinne sogleich angeben könnten, ob er zu ihr gehöre. Dahin gelangen wir nie.

Die regelmäßige Verknüpfung zwischen dem Zeichen, dessen Sinne und dessen Bedeutung ist der Art, daß dem Zeichen ein bestimmter Sinn und diesem wieder eine bestimmte Bedeutung entspricht, während zu einer Bedeutung (einem Gegenstande) nicht nur ein Zeichen zugehört. Derselbe Sinn hat in verschiedenen Sprachen, ja auch in derselben verschiedene Ausdrücke. Freilich kommen Ausnahmen von diesem regelmäßigen Verhalten vor. Gewiß sollte in einem vollkommenen Ganzen von Zeichen jedem

[2] Bei einem eigentlichen Eigennamen wie »Aristoteles« können freilich die Meinungen über den Sinn auseinandergehen. Man könnte z. B. als solchen annehmen: der Schüler Platos und Lehrer Alexanders des Großen. Wer dies tut, wird mit dem Satze »Aristoteles war aus Stagira gebürtig« einen andern Sinn verbinden als einer, der als Sinn dieses Namens annähme: der aus Stagira gebürtige Lehrer Alexanders des Großen. Solange nur die Bedeutung dieselbe bleibt, lassen sich diese Schwankungen des Sinnes ertragen, wiewohl auch sie in dem Lehrgebäude einer beweisenden Wissenschaft zu vermeiden sind und in einer vollkommenen Sprache nicht vorkommen dürften.

Ausdrucke ein bestimmter Sinn entsprechen; aber die Volkssprachen | erfüllen diese Forderung vielfach nicht, und man muß zufrieden sein, wenn nur in demselben Zusammenhange dasselbe Wort immer denselben Sinn hat. Vielleicht kann man zugeben, daß ein grammatisch richtig gebildeter Ausdruck, der für einen Eigennamen steht, immer einen Sinn habe. Aber ob dem Sinne nun auch eine Bedeutung entspreche, ist damit nicht gesagt. Die Worte »der von der Erde am weitesten entfernte Himmelskörper« haben einen Sinn; ob sie aber auch eine Bedeutung haben, ist sehr zweifelhaft. Der Ausdruck »die am wenigsten konvergente Reihe« hat einen Sinn; aber man beweist, daß er keine Bedeutung hat, da man zu jeder konvergenten Reihe eine weniger konvergente, aber immer noch konvergente finden kann. Dadurch also, daß man einen Sinn auffaßt, hat man noch nicht mit Sicherheit eine Bedeutung.

Wenn man in der gewöhnlichen Weise Worte gebraucht, so ist das, wovon man sprechen will, deren Bedeutung. Es kann aber auch vorkommen, daß man von den Worten selbst oder von ihrem Sinne reden will. Jenes geschieht z. B., wenn man die Worte eines andern in gerader Rede anführt. Die eigenen Worte bedeuten dann zunächst die Worte des andern, und erst diese haben die gewöhnliche Bedeutung. Wir haben dann Zeichen von Zeichen. In der Schrift schließt man in diesem Falle die Wortbilder in Anführungszeichen ein. Es darf also ein in Anführungszeichen stehendes Wortbild nicht in der gewöhnlichen Bedeutung genommen werden.

Wenn man von dem Sinne eines Ausdrucks ›A‹ reden will, so kann man dies einfach durch die Wendung »der Sinn des Ausdrucks ›A‹«. In der ungeraden Rede spricht man von dem Sinne z. B. der Rede eines andern. Es ist daraus klar, daß auch in dieser Redeweise die Worte nicht ihre gewöhnliche Bedeutung haben, sondern das bedeuten, was gewöhnlich ihr Sinn ist. Um einen kurzen Ausdruck zu haben, wollen wir sagen: die Wörter werden in der ungeraden Rede *ungerade* gebraucht oder haben ihre *ungerade* Bedeutung. Wir unterscheiden demnach die *gewöhnliche* Bedeutung eines Wortes von seiner *ungeraden* und seinen *gewöhnlichen* Sinn von seinem *ungeraden* Sinne. Die ungerade Bedeutung eines Wortes ist also sein gewöhnlicher Sinn. Solche Ausnahmen muß man immer im Auge behalten, wenn man die Verknüpfungsweise von Zeichen, Sinn und Bedeutung im einzelnen Falle richtig auffassen will.

Von der Bedeutung und dem Sinne eines Zeichens ist die mit ihm verknüpfte Vorstellung zu unterscheiden. Wenn die Bedeutung eines Zeichens ein sinnlich wahrnehmbarer Gegenstand ist, so ist meine Vorstellung davon ein aus Erinnerungen von Sinneseindrücken, die ich gehabt habe, und von Tätigkeiten, innern sowohl wie äußern, die ich ausgeübt habe, entstandenes inneres Bild[3]. Dieses ist oft mit Gefühlen getränkt; die Deutlichkeit seiner

[3] Wir können mit den Vorstellungen gleich die Anschauungen zusammennehmen, bei denen die Sinneseindrücke und die Tätigkeiten selbst an die Stelle der Spuren tre-

einzelnen Teile ist verschieden und schwankend. Nicht immer ist, auch bei demselben Menschen, dieselbe Vorstellung mit demselben Sinne verbunden. Die Vorstellung ist subjektiv: die Vorstellung des einen ist nicht die des andern. Damit sind von selbst mannigfache Unterschiede der mit demselben Sinne verknüpften Vorstellungen gegeben. Ein Maler, ein Reiter, ein Zoologe werden wahrscheinlich sehr verschiedene Vorstellungen mit dem Namen »Bucephalus« verbinden. Die Vorstellung unterscheidet sich dadurch wesentlich von dem Sinne eines Zeichens, welcher gemeinsames Eigentum von vielen sein kann und also nicht Teil oder Modus der Einzelseele ist; denn man wird wohl nicht leugnen können, daß die Menschheit einen gemeinsamen Schatz von Gedanken hat, den sie von einem Geschlechte auf das andere überträgt[4].

Während es demnach keinem Bedenken unterliegt, von dem Sinne schlechtweg zu sprechen, muß man bei der Vorstellung genau genommen hinzufügen, wem sie angehört und zu welcher Zeit. Man könnte vielleicht sagen: ebensogut, wie mit demselben Worte der eine diese, der andere jene Vorstellung verbindet, kann auch der eine diesen, der andere jenen Sinn damit verknüpfen. Doch besteht der Unterschied dann doch nur in der Weise dieser Verknüpfung. Das hindert nicht, daß beide denselben Sinn auffassen; | aber dieselbe Vorstellung können sie nicht haben. Si duo idem faciunt, non est idem. Wenn zwei sich dasselbe vorstellen, so hat jeder doch seine eigene Vorstellung. Es ist zwar zuweilen möglich, Unterschiede der Vorstellungen, ja der Empfindungen verschiedener Menschen festzustellen; aber eine genaue Vergleichung ist nicht möglich, weil wir diese Vorstellungen nicht in demselben Bewußtsein zusammen haben können.

Die Bedeutung eines Eigennamens ist der Gegenstand selbst, den wir damit bezeichnen; die Vorstellung, welche wir dabei haben, ist ganz subjektiv; dazwischen liegt der Sinn, der zwar nicht mehr subjektiv wie die Vorstellung, aber doch auch nicht der Gegenstand selbst ist. Folgendes Gleichnis ist vielleicht geeignet, diese Verhältnisse zu verdeutlichen. Jemand betrachtet den Mond durch ein Fernrohr. Ich vergleiche den Mond selbst mit der Bedeutung; er ist der Gegenstand der Beobachtung, die vermittelt wird durch das reelle Bild, welches vom Objektivglase im Innern des Fernrohrs entworfen wird, und durch das Netzhautbild des Betrachtenden. Jenes vergleiche ich mit dem Sinne, dieses mit der Vorstellung oder Anschauung. Das Bild im Fernrohre ist zwar nur einseitig; es ist abhängig vom Standorte; aber es ist doch objektiv, insofern es mehreren Beobachtern dienen kann. Es

ten, die sie in der Seele zurückgelassen haben. Der Unterschied ist für unsern Zweck unerheblich, zumal wohl immer neben den Empfindungen und Tätigkeiten Erinnerungen von solchen das Anschauungsbild vollenden helfen. Man kann unter Anschauung aber auch einen Gegenstand verstehen, sofern er sinnlich wahrnehmbar oder räumlich ist.

[4] Darum ist es unzweckmäßig, mit dem Worte »Vorstellung« so Grundverschiedenes zu bezeichnen.

ließe sich allenfalls einrichten, daß gleichzeitig mehrere es benutzen. Von den Netzhautbildern aber würde jeder doch sein eignes haben. Selbst eine geometrische Kongruenz würde wegen der verschiedenen Bildung der Augen kaum zu erreichen sein, ein wirkliches Zusammenfallen aber wäre ausgeschlossen. Dies Gleichnis ließe sich vielleicht noch weiter ausführen, indem man annähme, das Netzhautbild des A könnte dem B sichtbar gemacht werden; oder auch A selbst könnte in einem Spiegel sein eignes Netzhautbild sehen. Hiermit wäre vielleicht zu zeigen, wie eine Vorstellung zwar selbst zum Gegenstande genommen werden kann, als solche aber doch dem Betrachter nicht das ist, was sie unmittelbar dem Vorstellenden ist. Doch würde, dies zu verfolgen, wohl zu weit abführen.

Wir können nun drei Stufen der Verschiedenheit von Wörtern, Ausdrükken und ganzen Sätzen erkennen. Entweder betrifft der Unterschied höchstens die Vorstellungen, oder den Sinn aber nicht die Bedeutung, oder endlich auch die Bedeutung. In bezug auf | die erste Stufe ist zu bemerken, daß, wegen der unsichern Verbindung der Vorstellungen mit den Worten für den einen eine Verschiedenheit bestehen kann, die der andere nicht findet. Der Unterschied der Übersetzung von der Urschrift soll eigentlich die erste Stufe nicht überschreiten. Zu den hier noch möglichen Unterschieden gehören die Färbungen und Beleuchtungen, welche Dichtkunst Beredsamkeit dem Sinne zu geben suchen. Diese Färbungen und Beleuchtungen sind nicht objektiv, sondern jeder Hörer und Leser muß sie sich selbst nach den Winken des Dichters oder Redners hinzuschaffen. Ohne eine Verwandtschaft des menschlichen Vorstellens wäre freilich die Kunst nicht möglich; wieweit aber den Absichten des Dichters entsprochen wird, kann nie genau ermittelt werden.

Von den Vorstellungen und Anschauungen soll im folgenden nicht mehr die Rede sein; sie sind hier nur erwähnt worden, damit die Vorstellung, die ein Wort bei einem Hörer erweckt, nicht mit dessen Sinne oder dessen Bedeutung verwechselt werde.

Um einen kurzen und genauen Ausdruck möglich zu machen, mögen folgende Redewendungen festgesetzt werden:

Ein Eigenname (Wort, Zeichen, Zeichenverbindung, Ausdruck) drückt aus seinen Sinn, bedeutet oder bezeichnet seine Bedeutung. Wir drücken mit einem Zeichen dessen Sinn aus und bezeichnen mit ihm dessen Bedeutung.

Von idealistischer und skeptischer Seite ist vielleicht schon längst eingewendet worden: »du sprichst hier ohne weiteres von dem Monde als einem Gegenstande; aber woher weißt du, daß der Name ›der Mond‹ überhaupt eine Bedeutung hat, woher weißt du, daß überhaupt irgend etwas eine Bedeutung hat?« Ich antworte, daß es nicht unsere Absicht ist, von unserer Vorstellung des Mondes zu sprechen, und daß wir uns auch nicht mit dem Sinne begnügen, wenn wir ›der Mond‹ sagen; sondern wir setzen eine Bedeutung voraus. Es hieße den Sinn geradezu verfehlen, wenn man annehmen wollte, in dem Satze »der Mond ist kleiner als die Erde« sei von einer Vor-

stellung des Mondes die Rede. Wollte der Sprechende dies, so würde er die Wendung »meine Vorstellung vom Monde« gebrauchen. Nun können wir uns in jener Voraussetzung freilich irren, und solche Irrtümer sind auch vorgekommen. Die Frage aber, ob wir uns vielleicht immer darin irren, kann | hier unbeantwortet bleiben; es genügt zunächst, auf unsere Absicht beim Sprechen oder Denken hinzuweisen, um es zu rechtfertigen, von der Bedeutung eines Zeichens zu sprechen, wenn auch mit dem Vorbehalte: falls eine solche vorhanden ist.

Bisher sind Sinn und Bedeutung nur von solchen Ausdrücken, Wörtern, Zeichen betrachtet worden, welche wir Eigennamen genannt haben. Wir fragen nun nach Sinn und Bedeutung eines ganzen Behauptungssatzes. Ein solcher Satz enthält einen Gedanken[5]. Ist dieser Gedanke nun als dessen Sinn oder als dessen Bedeutung anzusehen? Nehmen wir einmal an, der Satz habe eine Bedeutung! Ersetzen wir nun in ihm ein Wort durch ein anderes von derselben Bedeutung, aber anderm Sinne, so kann dies auf die Bedeutung des Satzes keinen Einfluß haben. Nun sehen wir aber, daß der Gedanke sich in solchem Falle ändert; denn es ist z. B. der Gedanke des Satzes »der Morgenstern ist ein von der Sonne beleuchteter Körper« verschieden von dem des Satzes »der Abendstern ist ein von der Sonne beleuchteter Körper«. Jemand, der nicht wüßte, daß der Abendstern der Morgenstern ist, könnte den einen Gedanken für wahr, den andern für falsch halten. Der Gedanke kann also nicht die Bedeutung des Satzes sein, vielmehr werden wir ihn als den Sinn aufzufassen haben. Wie ist es nun aber mit der Bedeutung? Dürfen wir überhaupt danach fragen? Hat vielleicht ein Satz als Ganzes nur einen Sinn, aber keine Bedeutung? Man wird jedenfalls erwarten können, daß solche Sätze vorkommen, ebensogut, wie es Satzteile gibt, die wohl einen Sinn, aber keine Bedeutung haben. Und Sätze, welche Eigennamen ohne Bedeutung enthalten, werden von der Art sein. Der Satz »Odysseus wurde tief schlafend in Ithaka ans Land gesetzt« hat offenbar einen Sinn. Da es aber zweifelhaft ist, ob der darin vorkommende Name »Odysseus« eine Bedeutung habe, so ist es damit auch zweifelhaft, ob der ganze Satz eine habe. Aber sicher ist doch, daß jemand, der im Ernste den Satz für wahr oder für falsch hält, auch dem Namen »Odysseus« eine Bedeutung zuerkennt, nicht nur einen Sinn; denn der Bedeutung dieses | Namens wird ja das Prädikat zu- oder abgesprochen. Wer eine Bedeutung nicht anerkennt, der kann ihr ein Prädikat weder zu- noch absprechen. Nun wäre aber das Vordringen bis zur Bedeutung des Namens überflüssig; man könnte sich mit dem Sinne begnügen, wenn man beim Gedanken stehenbleiben wollte. Käme es nur auf den Sinn des Satzes, den Gedanken, an, so wäre es unnötig, sich um die Bedeutung eines Satzteils zu kümmern; für den Sinn des Satzes kann ja nur der Sinn, nicht die Bedeutung dieses Teiles in Betracht kommen. Der Gedanke bleibt derselbe, ob der Name »Odysseus« eine Bedeutung hat oder nicht. Daß wir uns überhaupt um die Bedeutung eines Satzteils bemühen, ist ein Zeichen dafür, daß wir auch für den Satz selbst eine Bedeutung im allgemeinen anerkennen und fordern. Der Gedanke verliert für uns an Wert, sobald wir erkennen, daß zu einem seiner Teile die Bedeutung fehlt. Wir sind also wohl berechtigt, uns nicht mit dem Sinne eines Satzes zu begnügen, sondern auch nach seiner Bedeutung zu fragen. Warum wollen wir denn aber, daß jeder Eigenname nicht nur einen Sinn, sondern auch eine Bedeutung habe? Warum genügt uns der Gedanke nicht? Weil und soweit es uns auf seinen Wahrheitswert ankommt. Nicht immer ist dies der Fall. Beim Anhören eines Epos z. B. fesseln uns neben dem Wohlklange der Sprache allein der Sinn der Sätze und die davon erweckten Vorstellungen und Gefühle. Mit der Frage nach der Wahrheit würden wir den Kunstgenuß verlassen und uns einer wissenschaftlichen Betrachtung zuwenden. Daher ist es uns auch gleichgültig, ob der Name »Odysseus« z. B. eine Bedeutung habe, solange wir das Gedicht als Kunstwerk aufnehmen[6]. Das Streben nach Wahrheit also ist es, was uns überall vom Sinne zur Bedeutung vorzudringen treibt.

Wir haben gesehn, daß zu einem Satze immer dann eine Bedeutung zu suchen ist, wenn es auf die Bedeutung der Bestandteile ankommt; und das ist immer dann und nur dann der Fall, wenn wir nach dem Wahrheitswerte fragen. |

So werden wir dahin gedrängt, den *Wahrheitswert* eines Satzes als seine Bedeutung anzuerkennen. Ich verstehe unter dem Wahrheitswerte eines Satzes den Umstand, daß er wahr oder daß er falsch ist. Weitere Wahrheitswerte gibt es nicht. Ich nenne der Kürze halber den einen das Wahre, den andern das Falsche. Jeder Behauptungssatz, in dem es auf die Bedeutung der Wörter ankommt, ist also als Eigenname aufzufassen, und zwar ist seine Bedeutung, falls sie vorhanden ist, entweder das Wahre oder das Falsche. Diese beiden Gegenstände werden von jedem, wenn auch nur stillschweigend, anerkannt, der überhaupt urteilt, der etwas für wahr hält, also auch vom Skeptiker. Die Bezeichnung der Wahrheitswerte als Gegenstände mag hier noch als willkürlicher Einfall und vielleicht als bloßes Spiel mit Worten erscheinen, aus dem man keine tiefgehenden Folgerungen ziehen dürfe. Was ich einen Gegenstand nenne, kann genauer nur im Zusammenhange mit Begriff und Beziehung erörtert werden. Das will ich einem andern Aufsatze vorbehalten. Aber so viel möchte doch schon hier klar sein, daß in jedem Urteile[7] — und sei es noch so selbstverständlich — schon der Schritt von der Stufe der Gedanken zur Stufe der Bedeutungen (des Objektiven) geschehen ist.

[5] Ich verstehe unter Gedanken nicht das subjektive Tun des Denkens, sondern dessen objektiven Inhalt, der fähig ist, gemeinsames Eigentum von vielen zu sein.

[6] Es wäre wünschenswert, für Zeichen, die nur einen Sinn haben sollen, einen besondern Ausdruck zu haben. Nennen wir solche etwa Bilder, so würden die Worte des Schauspielers auf der Bühne Bilder sein, ja, der Schauspieler selber wäre ein Bild.

[7] Ein Urteil ist mir nicht das bloße Fassen eines Gedankens, sondern die Anerkennung seiner Wahrheit.

Man könnte versucht sein, das Verhältnis des Gedankens zum Wahren nicht als das des Sinnes zur Bedeutung, sondern als das des Subjekts zum Prädikate anzusehen. Man kann ja geradezu sagen: »der Gedanke, daß 5 eine Primzahl ist, ist wahr«. Wenn man aber genauer zusieht, so bemerkt man, daß damit eigentlich nichts mehr gesagt ist als in dem einfachen Satze »5 ist eine Primzahl«. Die Behauptung der Wahrheit liegt in beiden Fällen in der Form des Behauptungssatzes, und da, wo diese nicht ihre gewöhnliche Kraft hat, z. B. im Munde eines Schauspielers auf der Bühne, enthält der Satz »der Gedanke, daß 5 eine Primzahl ist, ist wahr« eben auch nur einen Gedanken, und zwar denselben Gedanken wie das einfache »5 ist eine Primzahl«. Daraus ist zu entnehmen, daß das Verhältnis des Gedankens zum Wahren doch mit dem des Subjekts zum Prädikate nicht verglichen werden darf. | Subjekt und Prädikat sind ja (im logischen Sinne verstanden) Gedankenteile; sie stehen auf derselben Stufe für das Erkennen. Man gelangt durch die Zusammenfügung von Subjekt und Prädikat immer nur zu einem Gedanken, nie von einem Sinne zu dessen Bedeutung, nie von einem Gedanken zu dessen Wahrheitswerte. Man bewegt sich auf derselben Stufe, aber man schreitet nicht von einer Stufe zur nächsten vor. Ein Wahrheitswert kann nicht Teil eines Gedankens sein, sowenig wie etwa die Sonne, weil er kein Sinn ist, sondern ein Gegenstand.

Wenn unsere Vermutung richtig ist, daß die Bedeutung eines Satzes sein Wahrheitswert ist, so muß dieser unverändert bleiben, wenn ein Satzteil durch einen Ausdruck von derselben Bedeutung, aber anderm Sinne ersetzt wird. Und das ist in der Tat der Fall. Leibniz erklärt gradezu: »Eadem sunt, quae sibi mutuo substitui possunt, salva veritate.« Was sonst als der Wahrheitswert könnte auch gefunden werden, das ganz allgemein zu jedem Satze gehört, bei dem überhaupt die Bedeutung der Bestandteile in Betracht kommt, was bei einer Ersetzung der angegebenen Art unverändert bliebe?

Wenn nun der Wahrheitswert eines Satzes dessen Bedeutung ist, so haben einerseits alle wahren Sätze dieselbe Bedeutung, andrerseits alle falschen. Wir sehn daraus, daß in der Bedeutung des Satzes alles einzelne verwischt ist. Es kann uns also niemals auf die Bedeutung eines Satzes allein ankommen; aber auch der bloße Gedanke gibt keine Erkenntnis, sondern erst der Gedanke zusammen mit seiner Bedeutung, d. h. seinem Wahrheitswerte. Urteilen kann als Fortschreiten von einem Gedanken zu seinem Wahrheitswerte gefaßt werden. Freilich soll dies keine Definition sein. Das Urteilen ist eben etwas ganz Eigenartiges und Unvergleichliches. Man könnte auch sagen Urteilen sei Unterscheiden von Teilen innerhalb des Wahrheitswertes. Diese Unterscheidung geschieht durch Rückgang zum Gedanken. Jeder Sinn, der zu einem Wahrheitswerte gehört, würde einer eignen Weise der Zerlegung entsprechen. Das Wort »Teil« habe ich hier allerdings in besondrer Weise gebraucht. Ich habe nämlich das Verhältnis des Ganzen und des Teils vom Satze auf seine Bedeutung übertragen, indem ich die Bedeutung eines Wortes Teil der Bedeutung des Satzes genannt habe, wenn das Wort selbst |

Teil dieses Satzes ist, eine Redeweise, die freilich anfechtbar ist, weil bei der Bedeutung durch das Ganze und einen Teil der andere nicht bestimmt ist, und weil man bei Körpern das Wort Teil schon in anderm Sinne gebraucht. Es müßte ein eigener Ausdruck hierfür geschaffen werden.

Es soll nun die Vermutung, daß der Wahrheitswert eines Satzes dessen Bedeutung ist, weiter geprüft werden. Wir haben gefunden, daß der Wahrheitswert eines Satzes unberührt bleibt, wenn wir darin einen Ausdruck durch einen gleichbedeutenden ersetzen: wir haben aber dabei den Fall noch nicht betrachtet, daß der zu ersetzende Ausdruck selber ein Satz ist. Wenn nun unsere Ansicht richtig ist, so muß der Wahrheitswert eines Satzes, der einen andern als Teil enthält, unverändert bleiben, wenn wir für den Teilsatz einen andern einsetzen, dessen Wahrheitswert derselbe ist. Ausnahmen sind dann zu erwarten, wenn das Ganze oder der Teilsatz gerade oder ungerade Rede sind; denn, wie wir gesehen haben, ist die Bedeutung der Worte dann nicht die gewöhnliche. Ein Satz bedeutet in der geraden Rede wieder einen Satz und in der ungeraden einen Gedanken.

Wir werden so auf die Betrachtung der Nebensätze hingelenkt. Diese treten ja als Teile eines Satzgefüges auf, das vom logischen Gesichtspunkte aus gleichfalls als Satz, und zwar als Hauptsatz, erscheint. Aber es tritt uns hier die Frage entgegen, ob denn von den Nebensätzen gleichfalls gilt, daß ihre Bedeutung ein Wahrheitswert sei. Von der ungeraden Rede wissen wir ja schon das Gegenteil. Die Grammatiker sehen die Nebensätze als Vertreter von Satzteilen an und teilen sie danach ein in Nennsätze, Beisätze, Adverbsätze. Daraus könnte man die Vermutung schöpfen, daß die Bedeutung eines Nebensatzes nicht ein Wahrheitswert, sondern gleichartig sei der eines Nennworts oder Beiworts oder Adverbs, kurz eines Satzteils, der als Sinn keinen Gedanken, sondern nur einen Teil eines solchen hat. Nur eine eingehendere Untersuchung kann darüber Klarheit verschaffen. Wir werden uns dabei nicht streng an den grammatischen Leitfaden halten, sondern das zusammenfassen, was logisch gleichartig ist. Suchen wir zunächst solche Fälle auf, in denen der Sinn des Nebensatzes, wie wir eben vermuteten, kein selbständiger Gedanke ist. |

Zu den mit »daß« eingeleiteten abstrakten Nennsätzen gehört auch die ungerade Rede, von der wir gesehen haben, daß in ihr die Wörter ihre ungerade Bedeutung haben, welche mit dem übereinstimmt, was gewöhnlich ihr Sinn ist. In diesem Falle hat also der Nebensatz als Bedeutung einen Gedanken, keinen Wahrheitswert; als Sinn keinen Gedanken, sondern den Sinn der Worte »der Gedanke, daß . . .«, welcher nur Teil des Gedankens des ganzen Satzgefüges ist. Dies kommt vor nach »sagen«, »hören«, »meinen«, »überzeugt sein«, »schließen« und ähnlichen Wörtern[8]. Anders, und zwar

[8] In »A log, daß er den B gesehen habe« bedeutet der Nebensatz einen Gedanken, von dem erstens gesagt wird, daß A ihn als wahr behauptete, und zweitens, daß A von seiner Falschheit überzeugt war.

ziemlich verwickelt, liegt die Sache nach Wörtern wie »erkennen«, »wissen«, »wähnen«, was später zu betrachten sein wird.

Daß in unsern Fällen die Bedeutung des Nebensatzes in der Tat der Gedanke ist, sieht man auch daran, daß es für die Wahrheit des Ganzen gleichgültig ist, ob jener Gedanke wahr ist oder falsch. Man vergleiche z. B. die beiden Sätze: »Kopernikus glaubte, daß die Bahnen der Planeten Kreise seien« und »Kopernikus glaubte, daß der Schein der Sonnenbewegung durch die wirkliche Bewegung der Erde hervorgebracht werde.« Man kann hier unbeschadet der Wahrheit den einen Nebensatz für den andern einsetzen. Der Hauptsatz zusammen mit dem Nebensatze hat als Sinn nur einen einzigen Gedanken und die Wahrheit des Ganzen schließt weder die Wahrheit noch die Unwahrheit des Nebensatzes ein. In diesen Fällen ist es nicht erlaubt, in dem Nebensatze einen Ausdruck durch einen andern zu ersetzen, der dieselbe gewöhnliche Bedeutung hat, sondern nur durch einen solchen, welcher dieselbe ungerade Bedeutung, d. h. denselben gewöhnlichen Sinn hat. Wenn jemand schließen wollte: die Bedeutung eines Satzes ist nicht sein Wahrheitswert, »denn dann dürfte man ihn überall durch einen andern von demselben Wahrheitswerte ersetzen«, so würde er zuviel beweisen; ebensogut könnte man behaupten, die Bedeutung des Wortes »Morgenstern« sei nicht die Venus; denn man dürfe nicht überall für »Morgenstern« »Venus« sagen. Mit Recht kann man nur folgern, daß die Bedeutung des Satzes *nicht immer* sein Wahrheitswert ist, und daß »Morgenstern« nicht | immer den Planeten Venus bedeutet, nämlich dann nicht, wenn dies Wort seine ungerade Bedeutung hat. Ein solcher Ausnahmefall liegt in den eben betrachteten Nebensätzen vor, deren Bedeutung ein Gedanke ist.

Wenn man sagt »es scheint, daß . . .« so meint man »es scheint mir, daß . .«, oder »ich meine, daß . . .«. Wir haben also wieder den Fall. Ähnlich liegt die Sache bei Ausdrücken, wie »sich freuen«, »bedauern«, »billigen«, »tadeln«, »hoffen«, »fürchten«. Wenn Wellington sich gegen Ende der Schlacht bei Belle-Alliance freute, daß die Preußen kämen, so war der Grund seiner Freude eine Überzeugung. Wenn er sich getäuscht hätte, so würde er sich, solange sein Wahn dauerte, nicht minder gefreut haben, und bevor er die Überzeugung gewann, daß die Preußen kämen, konnte er sich nicht darüber freuen, obwohl sie in der Tat schon anrückten.

Wie eine Überzeugung oder ein Glaube Grund eines Gefühls ist, so kann sie auch Grund einer Überzeugung sein wie beim Schließen. In dem Satze: »Kolumbus schloß aus der Rundung der Erde, daß er nach Westen reisend Indien erreichen könne«, haben wir als Bedeutungen von Teilen zwei Gedanken, daß die Erde rund sei und daß Kolumbus nach Westen reisend Indien erreichen könne. Es kommt hier wieder nur darauf an, daß Kolumbus von dem einen und von dem andern überzeugt war und daß die eine Überzeugung Grund der andern war. Ob die Erde wirklich rund ist und Kolumbus nach Westen reisend wirklich Indien so, wie er dachte, erreichen konnte, ist für die Wahrheit unseres Satzes gleichgültig; aber nicht gleichgültig ist, ob wir für »die Erde« setzen »der Planet, welcher von einem Monde begleitet ist, dessen Durchmesser größer als der vierte Teil seines eignen ist«. Auch hier haben wir die ungerade Bedeutung der Worte.

Die Adverbsätze des Zwecks mit »damit« gehören auch hierher; denn offenbar ist der Zweck ein Gedanke; daher: ungerade Bedeutung der Worte, konjunktiv.

Der Nebensatz mit »daß« nach »befehlen«, »bitten«, »verbieten« würde in gerader Rede als Imperativ erscheinen. Ein solcher hat keine Bedeutung, sondern nur einen Sinn. Ein Befehl, eine Bitte sind zwar nicht Gedanken, aber sie stehn doch mit Gedanken auf derselben Stufe. Daher haben in den von »befehlen«, | »bitten« usw. abhängigen Nebensätzen die Worte ihre ungerade Bedeutung. Die Bedeutung eines solchen Satzes ist also nicht ein Wahrheitswert, sondern ein Befehl, eine Bitte u. dgl.

Ähnlich ist es bei der abhängigen Frage in Wendungen wie »zweifeln, ob«, »nicht wissen, was«. Daß auch hier die Wörter in ihrer ungeraden Bedeutung zu nehmen sind, ist leicht zu sehn. Die abhängigen Fragesätze mit »wer«, »was«, »wo«, »wann«, »wie«, »wodurch« usw. nähern sich zuweilen scheinbar sehr Adverbsätzen, in denen die Worte ihre gewöhnliche Bedeutung haben. Sprachlich unterscheiden sich diese Fälle durch den Modus des Verbs. Beim Konjunktiv haben wir abhängige Frage und ungerade Bedeutung der Worte, so daß ein Eigenname nicht allgemein durch einen andern desselben Gegenstandes ersetzt werden kann.

In den bisher betrachteten Fällen hatten die Worte im Nebensatze ihre ungerade Bedeutung und daraus wurde erklärlich, daß auch die Bedeutung des Nebensatzes selbst eine ungerade war; d. h. nicht ein Wahrheitswert, sondern ein Gedanke, ein Befehl, eine Bitte, eine Frage. Der Nebensatz konnte als Nennwort aufgefaßt werden, ja, man könnte sagen: als Eigenname jenes Gedankens, jenes Befehls usw., als welcher er in den Zusammenhang des Satzgefüges eintrat.

Wir kommen jetzt zu andern Nebensätzen, in denen die Worte zwar ihre gewöhnliche Bedeutung haben, ohne daß doch als Sinn ein Gedanke und als Bedeutung ein Wahrheitswert auftritt. Wie das möglich ist, wird am besten an Beispielen deutlich.

> »Der die elliptische Gestalt der Planetenbahnen entdeckte, starb im Elend.«

Wenn hier der Nebensatz als Sinn einen Gedanken hätte, so müßte es möglich sein, diesen auch in einem Hauptsatze auszudrücken. Aber dies geht nicht, weil das grammatische Subjekt »der« keinen selbständigen Sinn hat, sondern die Beziehungen auf den Nachsatz »starb im Elend« vermittelt. Daher ist auch der Sinn des Nebensatzes kein vollständiger Gedanke und seine Bedeutung kein Wahrheitswert, sondern Kepler. Man könnte einwenden, daß der Sinn des Ganzen doch als Teil einen Gedanken einschließe, nämlich daß es einen gab, der die elliptische Gestalt der Planetenbahnen

zuerst erkannte; denn wer das das Ganze für wahr | halte, könne diesen Teil nicht verneinen. Das letzte ist zweifellos; aber nur weil sonst der Nebensatz »der die elliptische Gestalt der Planetenbahnen entdeckte« keine Bedeutung hätte. Wenn man etwas behauptet, so ist immer die Voraussetzung selbstverständlich, daß die gebrauchten einfachen oder zusammengesetzten Eigennamen eine Bedeutung haben. Wenn man also behauptet, »Kepler starb im Elend«, so ist dabei vorausgesetzt, daß der Name »Kepler« etwas bezeichne; aber darum ist doch im Sinne des Satzes »Kepler starb im Elend« der Gedanke, daß der Name »Kepler« etwas bezeichne nicht enthalten. Wenn das der Fall wäre, müßte die Verneinung nicht lauten

»Kepler starb nicht im Elend«,

sondern

»Kepler starb nicht im Elend, oder der Name ›Kepler‹ ist bedeutungslos«

Daß der Name »Kepler« etwas bezeichne, ist vielmehr Voraussetzung ebenso für die Behauptung

»Kepler starb im Elend«

wie für die entgegengesetzte. Nun haben die Sprachen den Mangel, daß in ihnen Ausdrücke möglich sind, welche nach ihrer grammatischen Form bestimmt erscheinen, einen Gegenstand zu bezeichnen, diese ihre Bestimmung aber in besondern Fällen nicht erreichen, weil das von der Wahrheit eines Satzes abhängt. So hängt es von der Wahrheit des Satzes

»es gab einen, der die elliptische Gestalt der Planetenbahnen entdeckte«

ab, ob der Nebensatz

»der die elliptische Gestalt der Planetenbahnen entdeckte«

wirklich einen Gegenstand bezeichnet oder nur den Schein davon erweckt, in der Tat jedoch bedeutungslos ist. Und so kann es scheinen, als ob unser Nebensatz als Teil seines Sinnes den Gedanken enthalte, es habe einen gegeben, der die elliptische Gestalt der Planetenbahnen entdeckte. Wäre das richtig, so müßte die Verneinung lauten:

»der die elliptische Gestalt der Planetenbahnen zuerst erkannte, starb nicht im Elend, oder es gab keinen, der die elliptische Gestalt der Planetenbahnen entdeckte.« |

Dies liegt also an einer Unvollkommenheit der Sprache, von der übrigens auch die Zeichensprache der Analysis nicht ganz frei ist; auch da können Zeichenverbindungen vorkommen, die den Schein erwecken, als bedeuten sie etwas, die aber wenigstens bisher noch bedeutungslos sind, z. B.

divergente unendliche Reihen. Man kann dies vermeiden, z. B. durch die besondere Festsetzung, daß divergente unendliche Reihen die Zahl 0 bedeuten sollen. Von einer logisch vollkommenen Sprache (Begriffsschrift) ist zu verlangen, daß jeder Ausdruck, der aus schon eingeführten Zeichen in grammatisch richtiger Weise als Eigenname gebildet ist, auch in der Tat einen Gegenstand bezeichne und daß kein Zeichen als Eigenname neu eingeführt werde, ohne daß ihm eine Bedeutung gesichert sei. Man warnt in den Logiken vor der Vieldeutigkeit der Ausdrücke als einer Quelle von logischen Fehlern. Für mindestens ebenso angebracht halte ich die Warnung vor scheinbaren Eigennamen, die keine Bedeutung haben. Die Geschichte der Mathematik weiß von Irrtümern zu erzählen, die daraus entstanden sind. Der demagogische Mißbrauch liegt hierbei ebenso nahe, vielleicht näher als bei vieldeutigen Wörtern. »Der Wille des Volks« kann als Beispiel dazu dienen; denn, daß es wenigstens keine allgemein angenommene Bedeutung dieses Ausdrucks gibt, wird leicht festzustellen sein. Es ist also durchaus nicht belanglos, die Quelle dieser Irrtümer wenigstens für die Wissenschaft ein für allemal zu verstopfen. Dann werden solche Einwände wie der eben besprochene unmöglich, weil es dann nie von der Wahrheit eines Gedankens abhängen kann, ob ein Eigenname eine Bedeutung hat.

Wir können diesen Nennsätzen eine Art der Beisätze und Adverbsätze in der Betrachtung anschließen, welche logisch nahe mit ihnen verwandt sind.

Auch Beisätze dienen dazu, zusammengesetzte Eigennamen zu bilden, wenn sie auch nicht wie die Nennsätze allein dazu hinreichen. Diese Beisätze sind Beiwörtern gleichzuachten. Statt »die Quadratwurzel aus 4, die kleiner ist als 0« kann man auch sagen, »die negative Quadratwurzel aus 4«. Wir haben hier den Fall, daß aus einem Begriffsausdrucke ein zusammengesetzter Eigenname mit Hilfe des bestimmten Artikels im Singular gebildet wird, was jedenfalls dann erlaubt ist, wenn ein Gegenstand | und nur ein einziger unter den Begriff fällt[9]. Begriffsausdrücke können nun so gebildet werden, daß Merkmale durch Beisätze angegeben werden, wie in unserm Beispiele durch den Satz »die kleiner ist als 0«. Es ist einleuchtend, daß ein solcher Beisatz ebensowenig wie vorhin der Nennsatz als Sinn einen Gedanken noch als Bedeutung einen Wahrheitswert haben kann, sondern er hat als Sinn nur einen Teil eines Gedankens, der in manchen Fällen auch durch ein einzelnes Beiwort ausgedrückt werden kann. Auch hier wie bei jenen Nennsätzen fehlt das selbständige Subjekt und damit auch die Möglichkeit, den Sinn des Nebensatzes in einem selbständigen Hauptsatze wiederzugeben.

Örter, Zeitpunkte, Zeiträume sind, logisch betrachtet, Gegenstände; mithin ist die sprachliche Bezeichnung eines bestimmten Ortes, eines bestimmten Augenblicks oder Zeitraums als Eigenname aufzufassen. Adverbsätze

[9] Nach dem oben Bemerkten müßte einem solchen Ausdrucke eigentlich durch besondere Festsetzung immer eine Bedeutung gesichert werden, z. B. durch die Bestimmung, daß als seine Bedeutung die Zahl 0 zu gelten habe, wenn kein Gegenstand oder mehr als einer unter den Begriff fällt.

des Orts und der Zeit können nun zur Bildung eines solchen Eigennamens in ähnlicher Weise gebraucht werden, wie wir es eben von den Nenn- und Beisätzen gesehn haben. Ebenso können Ausdrücke für Begriffe, die Örter usw. unter sich fassen, gebildet werden. Auch hier ist zu bemerken, daß der Sinn dieser Nebensätze nicht in einem Hauptsatze wiedergegeben werden kann, weil ein wesentlicher Bestandteil, nämlich die Orts- oder Zeitbestimmung fehlt, die durch ein Relativpronomen oder ein Fügewort nur angedeutet ist[10].

Auch in den Bedingungssätzen ist meistens, wie wir es eben | bei Nenn-, Bei- und Adverbsätzen gesehn haben, ein unbestimmt andeutender Bestandteil anzuerkennen, dem im Nachsatze ein ebensolcher entspricht. Indem beide aufeinander hinweisen, verbinden sie beide Sätze zu einem Ganzen, das in der Regel nur einen Gedanken ausdrückt. In dem Satze

> »wenn eine Zahl kleiner als 1 und größer als 0 ist, so ist auch ihr Quadrat kleiner als 1 und größer als 0«

ist dieser Bestandteil »eine Zahl« im Bedingungssatze und »ihr« im Nachsatze. Eben durch diese Unbestimmtheit erhält der Sinn die Allgemeinheit, welche man von einem Gesetze erwartet. Eben dadurch wird aber auch bewirkt, daß der Bedingungssatz allein keinen vollständigen Gedanken als Sinn hat und mit dem Nachsatze zusammen einen Gedanken, und zwar nur einen einzigen, ausdrückt, dessen Teile nicht mehr Gedanken sind. Es ist im allgemeinen unrichtig, daß im hypothetischen Urteile zwei Urteile in Wechselbeziehung gesetzt werden. Wenn man so oder ähnlich sagt, gebraucht man das Wort »Urteil« in demselben Sinne, den ich mit dem Worte »Gedanke« verbunden habe, so daß ich dafür sagen würde: »in einem hypothetischen Gedanken werden zwei Gedanken in Wechselbeziehung gesetzt«. Dies könnte nur dann

[10] Es sind bei diesen Sätzen übrigens leicht verschiedene Auffassungen möglich. Den Sinn des Satzes »nachdem Schleswig-Holstein von Dänemark losgerissen war, entzweiten sich Preußen und Österreich« können wir auch wiedergeben in der Form »nach Losreißung Schleswig-Holsteins von Dänemark entzweiten sich Preußen und Österreich«. Bei dieser Fassung ist es wohl hinreichend deutlich, daß als Teil dieses Sinnes nicht der Gedanke aufzufassen ist, daß Schleswig-Holstein einmal von Dänemark losgerissen ist, sondern daß dies die notwendige Voraussetzung dafür ist, daß der Ausdruck »nach der Losreißung Schleswig-Holsteins von Dänemark« überhaupt eine Bedeutung habe. Es läßt sich freilich unser Satz auch so auffassen, daß damit gesagt sein soll, es sei einmal Schleswig-Holstein von Dänemark losgerissen worden. Dann haben wir einen Fall, der später zu betrachten sein wird. Versetzen wir uns, um den Unterschied klarer zu erkennen, in die Seele eines Chinesen, der bei seiner geringen Kenntnis europäischer Geschichte es für falsch hält, daß einmal Schleswig-Holstein von Dänemark losgerissen sei. Dieser wird unsern Satz, in der ersten Weise aufgefaßt, weder für wahr noch für falsch halten, sondern ihm jede Bedeutung absprechen, weil dem Nebensatze eine solche fehlen würde. Dieser würde nur scheinbar eine Zeitbestimmung geben. Wenn er unsern Satz dagegen in der zweiten Weise auffaßt, wird er in ihm einen Gedanken ausgedrückt finden, den er für falsch hielte, neben einem Teile, der für ihn bedeutungslos wäre.

wahr sein, wenn ein unbestimmt andeutender Bestandteil fehlte[11]; dann wäre aber auch keine Allgemeinheit vorhanden.

Wenn ein Zeitpunkt im Bedingungs- und Nachsatze unbestimmt anzudeuten ist, so geschieht es nicht selten nur durch das Tempus praesens des Verbs, das in diesem Falle nicht die Gegenwart mitbezeichnet. Diese grammatische Form ist dann im Haupt- und Nebensatze der unbestimmt andeutende Bestandteil. »Wenn sich | die Sonne im Wendekreise des Krebses befindet, haben wir auf der nördlichen Erdhälfte den längsten Tag«, ist ein Beispiel dafür. Auch hier ist es unmöglich, den Sinn des Nebensatzes in einem Hauptsatze auszudrücken, weil dieser Sinn kein vollständiger Gedanke ist; denn, wenn wir sagten: »die Sonne befindet sich im Wendekreise des Krebses«, so würden wir das auf unsere Gegenwart beziehen und damit den Sinn ändern. Ebensowenig ist der Sinn des Hauptsatzes ein Gedanke; erst das aus Haupt- und Nebensatz bestehende Ganze enthält einen solchen. Übrigens können auch mehrere gemeinsame Bestandteile im Bedingungs- und Nachsatze unbestimmt angedeutet werden.

Es ist einleuchtend, daß Nennsätze mit »wer«, »was« und Adverbsätze mit »wo«, »wann«, »wo immer«, »wann immer« vielfach als Bedingungssätze dem Sinne nach aufzufassen sind, z. B. »Wer Pech angreift, besudelt sich«.

Auch Beisätze können Bedingungssätze vertreten. So können wir den Sinn unseres vorhin angeführten Satzes auch in der Form »das Quadrat einer Zahl, die kleiner als 1 und größer als 0 ist, ist kleiner als 1 und größer als 0« ausdrücken.

Ganz anders wird die Sache, wenn der gemeinsame Bestandteil von Hauptsatz und Nebensatz durch einen Eigennamen bezeichnet wird. In dem Satze:

> »Napoleon, der die Gefahr für seine rechte Flanke erkannte, führte selbst seine Garden gegen die feindliche Stellung«

sind die beiden Gedanken ausgedrückt:

1. Napoleon erkannte die Gefahr für seine rechte Flanke;
2. Napoleon führte selbst seine Garden gegen die feindliche Stellung.

Wann und wo dies geschah, kann zwar nur aus dem Zusammenhange erkannt werden, ist aber als dadurch bestimmt anzusehen. Wenn wir unsern ganzen Satz als Behauptung aussprechen, so behaupten wir damit zugleich die beiden Teilsätze. Wenn einer dieser Teilsätze falsch ist, so ist damit das Ganze falsch. Hier haben wir den Fall, daß der Nebensatz für sich allein als Sinn einen vollständigen Gedanken hat (wenn wir ihn durch Zeit- und Ortsangabe ergänzen). Die Bedeutung des Nebensatzes ist demnach ein Wahrheitswert. Wir können also erwarten, daß er sich unbeschadet der Wahrheit

[11] Zuweilen fehlt eine ausdrückliche sprachliche Andeutung und muß dem ganzen Zusammenhange entnommen werden.

des Ganzen durch einen Satz von dem- | selben Wahrheitswerte ersetzen lasse. Dies ist auch der Fall; nur muß beachtet werden, daß sein Subjekt »Napoleon« sein muß aus einem rein grammatischen Grunde, weil er nur dann in die Form eines zu »Napoleon« gehörenden Beisatzes gebracht werden kann. Sieht man aber von der Forderung ab, ihn in dieser Form zu sehn, und läßt man auch die Anreihung mit »und« zu, so fällt diese Beschränkung hinweg.

Auch in Nebensätzen mit »obgleich« werden vollständige Gedanken ausgedrückt. Dieses Fügewort hat eigentlich keinen Sinn und verändert auch den Sinn des Satzes nicht, sondern beleuchtet ihn nur in eigentümlicher Weise[12]. Wir könnten zwar unbeschadet der Wahrheit des Ganzen den Konzessivsatz durch einen andern desselben Wahrheitswertes ersetzen; aber die Beleuchtung würde dann leicht unpassend erscheinen, wie wenn man ein Lied traurigen Inhalts nach einer lustigen Weise singen wollte.

In den letzten Fällen schloß die Wahrheit des Ganzen die Wahrheit der Teilsätze ein. Anders ist es, wenn ein Bedingungssatz einen vollständigen Gedanken ausdrückt, indem er statt des nur andeutenden Bestandteils einen Eigennamen enthält oder etwas, was dem gleichzuachten ist. In dem Satze

> »wenn jetzt die Sonne schon aufgegangen ist, ist der Himmel stark bewölkt«

ist die Zeit die Gegenwart, also bestimmt. Auch der Ort ist als bestimmt zu denken. Hier kann man sagen, daß eine Beziehung zwischen den Wahrheitswerten des Bedingungs- und Folgesatzes gesetzt sei, nämlich die, daß der Fall nicht stattfinde, wo der Bedingungssatz das Wahre und der Nachsatz das Falsche bedeute. Danach ist unser Satz wahr, sowohl wenn jetzt die Sonne noch nicht aufgegangen ist, sei nun der Himmel stark bewölkt oder nicht, als auch wenn die Sonne schon aufgegangen ist und der Himmel stark bewölkt ist. Da es hierbei nur auf die Wahrheitswerte ankommt, so kann man jeden der Teilsätze durch einen andern von gleichem Wahrheitswerte ersetzen, ohne den Wahrheitswert des Ganzen zu ändern. Freilich würde auch hier die Beleuchtung meistens unpassend werden; der Gedanke würde leicht abgeschmackt | erscheinen; aber das hat mit seinem Wahrheitswerte nichts zu tun. Man muß dabei immer beachten, daß Nebengedanken mit anklingen, die aber nicht eigentlich ausgedrückt sind und darum in den Sinn des Satzes nicht eingerechnet werden dürfen, auf deren Wahrheitswert es also nicht ankommen kann[13].

Damit möchten die einfachen Fälle besprochen sein. Werfen wir hier einen Blick auf das Erkannte zurück!

[12] Ähnliches haben wir bei »aber«, »doch«.

[13] Man könnte den Gedanken unseres Satzes auch so ausdrücken: »entweder ist jetzt die Sonne noch nicht aufgegangen, oder der Himmel ist stark bewölkt«, woraus zu ersehen, wie diese Art der Satzverbindung aufzufassen ist.

Der Nebensatz hat meistens als Sinn keinen Gedanken, sondern nur einen Teil eines solchen und folglich als Bedeutung keinen Wahrheitswert. Dies hat entweder darin seinen Grund, daß im Nebensatze die Wörter ihre ungerade Bedeutung haben, so daß die Bedeutung, nicht der Sinn des Nebensatzes ein Gedanke ist, oder darin, daß der Nebensatz wegen eines darin nur unbestimmt andeutenden Bestandteils unvollständig ist, so daß er erst mit dem Hauptsatze zusammen einen Gedanken ausdrückt. Es kommen aber auch Fälle vor, wo der Sinn des Nebensatzes ein vollständiger Gedanke ist, und dann kann er unbeschadet der Wahrheit des Ganzen durch einen andern von denselben Wahrheitswerte ersetzt werden, soweit nicht grammatische Hindernisse vorliegen.

Wenn man alle aufstoßenden Nebensätze hierauf ansieht, so wird man bald solche treffen, die nicht recht in diese Fächer passen wollen. Der Grund davon wird, soviel ich sehe, darin liegen, daß diese Nebensätze keinen so einfachen Sinn haben. Fast immer, scheint es, verbinden wir mit einem Hauptgedanken, den wir aussprechen, Nebengedanken, die auch der Hörer, obwohl sie nicht ausgedrückt werden, mit unsern Worten verknüpft nach psychologischen Gesetzen. Und weil sie so von selbst mit unsern Worten verbunden erscheinen, fast wie der Hauptgedanke selbst, so wollen wir dann auch wohl einen solchen Nebengedanken mit ausdrücken. Dadurch wird der Sinn des Satzes reicher, und es kann wohl geschehn, daß wir mehr einfache Gedanken als Sätze haben. In manchen Fällen muß der Satz so verstanden werden, in andern kann es zweifelhaft sein, ob der Nebengedanke mit zum Sinne des Satzes gehört oder | ihn nur begleitet[14]. So könnte man vielleicht finden, daß in dem Satze

> »Napoleon, der die Gefahr für seine rechte Flanke erkannte,
> führte selbst seine Garden gegen die feindliche Stellung«

nicht nur die beiden oben angegebenen Gedanken ausgedrückt wären, sondern auch der, daß die Erkenntnis der Gefahr der Grund war, weshalb er die Garden gegen die feindliche Stellung führte. Man kann in der Tat zweifelhaft sein, ob dieser Gedanke nur leicht angeregt oder ob er wirklich ausgedrückt wird. Man lege sich die Frage vor, ob unser Satz falsch wäre, wenn Napoleons Entschluß schon vor der Wahrnehmung der Gefahr gefaßt wäre. Könnte unser Satz trotzdem wahr sein, so wäre unser Nebengedanke nicht als Teil des Sinnes unsers Satzes aufzufassen. Wahrscheinlich wird man sich dafür entscheiden. Im andern Falle würde die Sachlage recht verwickelt: wir hätten dann mehr einfache Gedanken als Sätze. Wenn wir nun auch den Satz

> »Napoleon erkannte die Gefahr für seine rechte Flanke«

durch einen andern desselben Wahrheitswertes ersetzten, z. B. durch

[14] Für die Frage, ob eine Behauptung eine Lüge, ein Eid ein Meineid sei, kann dies von Wichtigkeit werden.

»Napoleon war schon über 45 Jahre alt«,

so würde damit nicht nur unser erster, sondern auch unser dritter Gedanke geändert, und damit könnte auch dessen Wahrheitswert ein anderer werden — dann nämlich, wenn sein Alter nicht Grund des Entschlusses war, die Garden gegen den Feind zu führen. Hieraus ist zu sehn, weshalb in solchen Fällen nicht immer Sätze von demselben Wahrheitswert füreinander eintreten können. Der Satz drückt dann eben vermöge seiner Verbindung mit einem andern mehr aus als für sich allein.

Betrachten wir nun Fälle, wo solches regelmäßig vorkommt. In dem Satze

»Bebel wähnt, daß durch die Rückgabe Elsaß-Lothringens Frankreichs Rachegelüste beschwichtigt werden können«

sind zwei Gedanken ausgedrückt, von denen aber nicht der eine dem Haupt-, der andere dem Nebensatze angehört, nämlich

1. Bebel glaubt, daß durch die Rückgabe Elsaß-Lothringens Frankreichs Rachegelüste beschwichtigt werden können;
2. durch die Rückgabe Elsaß-Lothringens können Frankreichs Rachegelüste nicht beschwichtigt werden.

In dem Ausdrucke des ersten Gedankens haben die Worte des Nebensatzes ihre ungerade Bedeutung, während dieselben Worte im Ausdrucke des zweiten Gedankens ihre gewöhnliche Bedeutung haben. Wir sehn daraus, daß der Nebensatz in unserm ursprünglichen Satzgefüge eigentlich doppelt zu nehmen ist mit verschiedenen Bedeutungen, von denen die eine ein Gedanke, die andere ein Wahrheitswert ist. Weil nun der Wahrheitswert nicht die ganze Bedeutung des Nebensatzes ist, können wir diesen nicht einfach durch einen andern desselben Wahrheitswertes ersetzen. Ähnliches haben wir bei Ausdrücken wie »wissen«, »erkennen«, »es ist bekannt«.

Mit einem Nebensatze des Grundes und dem zugehörigen Hauptsatze drücken wir mehrere Gedanken aus, die aber nicht den Sätzen einzeln entsprechen. Der Satz

»weil das Eis spezifisch leichter als Wasser ist, schwimmt es auf dem Wasser«

haben wir

1. das Eis ist spezifisch leichter als Wasser;
2. wenn etwas spezifisch leichter als Wasser ist, so schwimmt es auf dem Wasser;
3. das Eis schwimmt auf dem Wasser.

Der dritte Gedanke brauchte allenfalls nicht ausdrücklich aufgeführt zu werden als in den ersten beiden enthalten. Dagegen würden weder der erste und dritte noch der zweite und dritte zusammen den Sinn unseres Satzes ausmachen. Man sieht nun, daß in unserm Nebensatze

»weil das Eis spezifisch leichter als Wasser ist«

sowohl unser erster Gedanke als auch ein Teil unsers zweiten ausgedrückt ist. Daher kommt es, daß wir unsern Nebensatz nicht einfach durch einen andern desselben Wahrheitswertes ersetzen können; denn dadurch würde auch unser zweiter Gedanke geändert, und davon könnte leicht auch dessen Wahrheitswert berührt werden.

Ähnlich ist die Sache in dem Satze

»wenn Eisen spezifisch leichter als Wasser wäre, so würde es auf dem Wasser schwimmen«.

Wir haben hier die beiden Gedanken, daß Eisen nicht spezifisch leichter ist als Wasser und daß etwas auf dem Wasser schwimmt, wenn es spezifisch leichter als Wasser ist. Der Nebensatz drückt wieder den einen und einen Teil des andern Gedankens aus.

Wenn wir den früher betrachteten Satz

»nachdem Schleswig-Holstein von Dänemark losgerissen war, entzweiten sich Preußen und Österreich«

so auffassen, daß darin der Gedanke ausgedrückt ist, es sei einmal Schleswig-Holstein von Dänemark losgerissen worden, so haben wir erstens diesen Gedanken, zweitens den Gedanken, daß zu einer Zeit, die durch den Nebensatz näher bestimmt ist, Preußen und Österreich sich entzweiten. Auch hier drückt dann der Nebensatz nicht nur einen Gedanken, sondern auch einen Teil eines andern aus. Daher darf man ihn nicht allgemein durch einen andern desselben Wahrheitswertes ersetzen.

Es ist schwer, alle in der Sprache gegebenen Möglichkeiten zu erschöpfen; aber ich hoffe doch im wesentlichen die Gründe aufgefunden zu haben, warum nicht immer unbeschadet der Wahrheit des ganzen Satzgefüges ein Nebensatz durch einen andern desselben Wahrheitswertes vertreten werden kann. Diese sind

1. daß der Nebensatz keinen Wahrheitswert bedeutet, indem er nur einen Teil eines Gedankens ausdrückt;
2. daß der Nebensatz zwar einen Wahrheitswert bedeutet, aber sich nicht darauf beschränkt, indem sein Sinn außer einem Gedanken auch noch einen Teil eines andern Gedankens umfaßt.

Der erste Fall tritt ein

a) bei der ungeraden Bedeutung der Worte,

b) wenn ein Teil des Satzes nur unbestimmt andeutet, statt ein Eigenname zu sein.

Im zweiten Falle kann der Nebensatz doppelt zu nehmen sein, nämlich einmal in gewöhnlicher Bedeutung, das andre Mal in ungerader Bedeutung; oder es kann der Sinn eines Teiles des Nebensatzes zugleich Bestandteil eines andern Gedankens sein, der mit dem unmittelbar im Nebensatze ausgedrückten zusammen den ganzen Sinn des Haupt- und Nebensatzes ausmacht.

Hieraus geht wohl mit hinreichender Wahrscheinlichkeit hervor, daß die Fälle, wo ein Nebensatz nicht durch einen andern desselben Wahrheitswertes ersetzbar ist, nichts gegen unsere Ansicht beweisen, | der Wahrheitswert sei die Bedeutung des Satzes, dessen Sinn ein Gedanke ist.

Kehren wir nun zu unserm Ausgangspunkte zurück!

Wenn wir den Erkenntniswert von »$a = a$« und »$a = b$« im allgemeinen verschieden fanden, so erklärt sich das dadurch, daß für den Erkenntniswert der Sinn des Satzes, nämlich der in ihm ausgedrückte Gedanke, nicht minder in Betracht kommt als seine Bedeutung, das ist sein Wahrheitswert. Wenn nun $a = b$ ist, so ist zwar die Bedeutung von »b« dieselbe wie die von »a« und also auch der Wahrheitswert von »$a = b$« derselbe wie von »$a = a$«. Trotzdem kann der Sinn von »b« von dem Sinne von »a« verschieden sein, und mithin auch der in »$a = b$« ausgedrückte Gedanke verschieden von dem »$a = a$« ausgedrückten sein; dann haben beide Sätze auch nicht denselben Erkenntniswert. Wenn wir wie oben unter «Urteil« verstehn den Fortschritt vom Gedanken zu dessen Wahrheitswerte, so werden wir auch sagen, daß die Urteile verschieden sind.

LECTURE XI

WHEN we originally contrasted the performative with the constative utterance we said that

(1) the performative should be doing something as opposed to just saying something; and
(2) the performative is happy or unhappy as opposed to true or false.

Were these distinctions really sound? Our subsequent discussion of doing and saying certainly seems to point to the conclusion that whenever I 'say' anything (except perhaps a mere exclamation like 'damn' or 'ouch') I shall be performing both locutionary and illocutionary acts, and these two kinds of acts seem to be the very things which we tried to use, under the names of 'doing' and 'saying', as a means of distinguishing performatives from constatives. If we are in general always doing both things, how can our distinction survive?

Let us first reconsider the contrast from the side of constative utterances: Of these, we were content to refer to 'statements' as the typical or paradigm case. Would it be correct to say that when we state something

(1) we are doing something as well as and distinct from just saying something, and

(2) our utterance is liable to be happy or unhappy (as well as, if you will, true or false)?

(1) Surely to state is every bit as much to perform an illocutionary act as, say, to warn or to pronounce. Of course it is not to perform an act in some specially physical way, other than in so far as it involves, when verbal, the making of movements of vocal organs; but then nor, as we have seen, is to warn, to protest, to promise or to name. 'Stating' seems to meet all the criteria we had for distinguishing the illocutionary act. Consider such an unexceptionable remark as the following:

> In saying that it was raining, I was not betting or arguing or warning: I was simply stating it as a fact.

Here 'stating' is put absolutely on a level with arguing, betting, and warning. Or again:

> In saying that it was leading to unemployment, I was not warning or protesting: I was simply stating the facts.

Or to take a different type of test also used earlier, surely

> I state that he did not do it

is exactly on a level with

> I argue that he did not do it,
> I suggest that he did not do it,
> I bet that he did not do it, &c.

If I simply use the primary or non-explicit form of utterance:

> He did not do it

we may make explicit what we were doing in saying this, or specify the illocutionary force of the utterance, equally by saying any of the above three (or more) things.

Moreover, although the utterance 'He did not do it' is often issued as a statement, and is then undoubtedly true or false (*this* is if anything is), it does not seem possible to say that it differs from 'I state that he did not do it' in this respect. If someone says 'I state that he did not do it', we investigate the truth of his statement in just the same way as if he had said 'He did not do it' *simpliciter*, when we took that to be, as we naturally often should, a statement. That is, to say 'I state that he did not' is to make the very same statement as to say 'He did not': it is not to make a different statement about what 'I' state (except in exceptional cases: the historic and habitual present, &c.). As notoriously, when I say even 'I think he did it' someone is being rude if he says 'That's a statement about you': and this *might* conceivably be about myself, whereas 'I state that he did it' could not. So that there is no necessary conflict between

(*a*) our issuing the utterance being the doing of something,
(*b*) our utterance being true or false.

For that matter compare, for example, 'I warn you that

it is going to charge', where likewise it is both a warning and true or false that it is going to charge; and that comes in in appraising the warning just as much as, though not quite in the same way as, in appraising the statement.

On mere inspection, 'I state that' does not appear to differ in any essential way from 'I maintain that' (to say which is to maintain that), 'I inform you that', 'I testify that', &c. Possibly some 'essential' differences may yet be established between such verbs: but nothing has been done towards this yet.

(2) Moreover, if we think of the second alleged contrast, according to which performatives are happy or unhappy and statements true or false, again from the side of supposed constative utterances, notably statements, we find that statements *are* liable to every kind of infelicity to which performatives are liable. Let us look back again and consider whether statements are not liable to precisely the same disabilities as, say, warnings by way of what we called 'infelicities'—that is various disabilities which make an utterance unhappy without, however, making it true or false.

We have already noted that sense in which saying, as equivalent to stating, 'The cat is on the mat' implies that I believe that the cat is on the mat. This is parallel to the sense—is the same sense—as that in which 'I promise to be there' implies that I intend to be there and that I believe I shall be able to be there. So the statement is liable to the *insincerity* form of infelicity; and even to the

breach form of infelicity in this sense, that saying or stating that the cat is on the mat commits me to saying or stating 'The mat is underneath the cat' just as much as the performative 'I define X as Y' (in the *fiat* sense say) commits me to using those terms in special ways in future discourse, and we can see how this is connected with such acts as promising. This means that statements can give rise to infelicities of our two Γ kinds.

Now what about infelicities of the A and B kinds, which rendered the act—warning, undertaking, &c.—null and void?: can a thing that looks like a statement be null and void just as much as a putative contract? The answer seems to be Yes, importantly. The first cases are A. 1 and A. 2, where there is no convention (or not an accepted convention) or where the circumstances are not appropriate for its invocation by the speaker. Many infelicities of just this type do infect statements.

We have already noticed the case of a putative statement *presupposing* (as it is called) the existence of that which it refers to; if no such thing exists, 'the statement' is not about anything. Now some say that in these circumstances, if, for example, someone asserts that the present King of France is bald, 'the question whether he is bald does not arise'; but it is better to say that the putative statement is null and void, exactly as when I say that I sell you something but it is not mine or (having been burnt) is not any longer in existence. Contracts often are void because the objects they are about do not exist, which involves a breakdown of reference.

But it is important to notice also that 'statements' too are liable to infelicity of this kind in other ways also parallel to contracts, promises, warnings, &c. Just as we often say, for example, 'You cannot order me', in the sense 'You have not the right to order me', which is equivalent to saying that you are not in the appropriate position to do so: so often there are things you cannot state—have no right to state—are not in a position to state. You *cannot* now state how many people there are in the next room; if you say 'There are fifty people in the next room', I can only regard you as guessing or conjecturing (just as sometimes you are not ordering me, which would be inconceivable, but possibly asking me to rather impolitely, so here you are 'hazarding a guess' rather oddly). Here there is something you might, in other circumstances, be in a position to state; but what about statements about other persons' feelings or about the future? Is a forecast or even a prediction about, say, persons' behaviour really a statement? It is important to take the speech-situation as a whole.

Just as sometimes we cannot appoint but only confirm an appointment already made, so sometimes we cannot state but only confirm a statement already made.

Putative statements are also liable to infelicities of type B, flaws, and hitches. Somebody 'says something he did not really mean'—uses the wrong word—says 'the cat is on the mat' when he meant to say 'bat'. Other similar trivialities arise—or rather not entirely trivialities; because it is possible to discuss such utterances

entirely in terms of meaning as equivalent to sense and reference and so get confused about them, though they are really easy to understand.

Once we realize that what we have to study is *not* the sentence but the issuing of an utterance in a speech situation, there can hardly be any longer a possibility of not seeing that stating is performing an act. Moreover, comparing stating to what we have said about the illocutionary act, it is an act to which, just as much as to other illocutionary acts, it is essential to 'secure uptake': the doubt about whether I stated something if it was not heard or understood is just the same as the doubt about whether I warned *sotto voce* or protested if someone did not take it as a protest, &c. And statements do 'take effect' just as much as 'namings', say: if I have stated something, then that commits me to other statements: other statements made by me will be in order or out of order. Also some statements or remarks made by you will be henceforward contradicting me or not contradicting me, rebutting me or not rebutting me, and so forth. If perhaps a statement does not invite a response, that is not essential to all illocutionary acts anyway. And certainly in stating we are or may be performing perlocutionary acts of all kinds.

The most that might be argued, and with some plausibility, is that there is no perlocutionary *object* specifically associated with stating, as there is with informing, arguing, &c.; and this comparative purity may be one reason why we give 'statements' a certain special

position. But this certainly would not justify giving, say, 'descriptions', if properly used, a similar priority, and it is in any case true of many illocutionary acts.

However, looking at the matter from the side of performatives, we may still feel that they lack something which statements have, even if, as we have shown, the converse is not so. Performatives are, of course, incidentally saying something as well as doing something, but we may feel that they are not essentially true or false as statements are. We may feel that there is here a dimension in which we judge, assess, or appraise the constative utterance (granting as a preliminary that it is felicitous) which does not arise with non-constative or performative utterances. Let us agree that all these circumstances of situation have to be in order for me to have succeeded in stating something, yet when I have, *the* question arises, was what I stated true or false? And this we feel, speaking in popular terms, is now the question of whether the statement 'corresponds with the facts'. With this I agree: attempts to say that the use of the expression 'is true' is equivalent to endorsing or the like are no good. So we have here a new dimension of criticism of the accomplished statement.

But now

(1) doesn't just such a similar objective assessment of the accomplished utterance arise, at least in many cases, with other utterances which seem typically performative; and

(2) is not this account of statements a little oversimplified?

First, there is an obvious slide towards truth or falsity in the case of, for example, verdictives, such as estimating, finding, and pronouncing. Thus we may:

estimate	rightly or wrongly	for example, that it is half past two,
find	correctly or incorrectly	for example, that he is guilty,
pronounce	correctly or incorrectly	for example, that the batsman is out.

We shall not say 'truly' in the case of verdictives, but we shall certainly address ourselves to the same question; and such adverbs as 'rightly', 'wrongly', 'correctly', and 'incorrectly' are used with statements too.

Or again there is a parallel between inferring and arguing soundly or validly and stating truly. It is not just a question of whether he did argue or infer but also of whether he had a right to, and did he succeed. Warning and advising may be done correctly or incorrectly, well or badly. Similar considerations arise about praise, blame, and congratulation. Blame is not in order, if, say, you have done the same thing yourself; and the question always arises whether the praise, blame, or congratulation was merited or unmerited: it is not enough to say that you have blamed him and there's an end on't—still one act is, with reason, preferred to another. The question whether praise and blame are merited is quite different

from the question whether they are opportune, and the same distinction can be made in the case of advice. It is a different thing to say that advice is good or bad from saying that it is opportune or inopportune, though the timing of advice is more important to its goodness than the timing of blame is to its being merited.

Can we be sure that stating truly is a different *class* of assessment from arguing soundly, advising well, judging fairly, and blaming justifiably? Do these not have something to do in complicated ways with facts? The same is true also of exercitives such as naming, appointing, bequeathing, and betting. Facts come in as well as our knowledge or opinion about facts.

Well, of course, attempts are constantly made to effect this distinction. The soundness of arguments (if they are not deductive arguments which are 'valid') and the meritedness of blame are not objective matters, it is alleged; or in warning, we are told, we should distinguish the 'statement' that the bull is about to charge from the warning itself. But consider also for a moment whether the question of truth or falsity is so very objective. We ask: 'Is it a *fair* statement?', and are the good reasons and good evidence for stating and saying so very different from the good reasons and evidence for performative acts like arguing, warning, and judging? Is the constative, then, always true or false? When a constative is confronted with the facts, we in fact appraise it in ways involving the employment of a vast array of terms which overlap with those that we use in the appraisal of

performatives. In real life, as opposed to the simple situations envisaged in logical theory, one cannot always answer in a simple manner whether it is true or false.

Suppose that we confront 'France is hexagonal' with the facts, in this case, I suppose, with France, is it true or false? Well, if you like, up to a point; of course I can see what you mean by saying that it is true for certain intents and purposes. It is good enough for a top-ranking general, perhaps, but not for a geographer. 'Naturally it is pretty rough', we should say, 'and pretty good as a pretty rough statement'. But then someone says: 'But is it true or is it false? I don't mind whether it is rough or not; of course it's rough, but it has to be true or false—it's a statement, isn't it?' How can one answer this question, whether it is true or false that France is hexagonal? It is just rough, and that is the right and final answer to the question of the relation of 'France is hexagonal' to France. It is a rough description; it is not a true or a false one.

Again, in the case of stating truly or falsely, just as much as in the case of advising well or badly, the intents and purposes of the utterance and its context are important; what is judged true in a school book may not be so judged in a work of historical research. Consider the constative, 'Lord Raglan won the battle of Alma', remembering that Alma was a soldier's battle if ever there was one and that Lord Raglan's orders were never transmitted to some of his subordinates. Did Lord Raglan then win the battle of Alma or did he not? Of

course in some contexts, perhaps in a school book, it is perfectly justifiable to say so—it is something of an exaggeration, maybe, and there would be no question of giving Raglan a medal for it. As 'France is hexagonal' is rough, so 'Lord Raglan won the battle of Alma' is exaggerated and suitable to some contexts and not to others; it would be pointless to insist on its truth or falsity.

Thirdly, let us consider the question whether it is true that all snow geese migrate to Labrador, given that perhaps one maimed one sometimes fails when migrating to get quite the whole way. Faced with such problems, many have claimed, with much justice, that utterances such as those beginning 'All . . .' are prescriptive definitions or advice to adopt a rule. But what rule? This idea arises partly through not understanding the reference of such statements, which is limited to the known; we cannot quite make the simple statement that the truth of statements depends on facts as distinct from knowledge of facts. Suppose that before Australia is discovered *X* says 'All swans are white'. If you later find a black swan in Australia, is *X* refuted? Is his statement false now? Not necessarily: he will take it back but he could say 'I wasn't talking about swans absolutely everywhere; for example, I was not making a statement about possible swans on Mars'. Reference depends on knowledge at the time of utterance.

The truth or falsity of statements is affected by what they leave out or put in and by their being misleading,

and so on. Thus, for example, descriptions, which are said to be true or false or, if you like, are 'statements', are surely liable to these criticisms, since they are selective and uttered for a purpose. It is essential to realize that 'true' and 'false', like 'free' and 'unfree', do not stand for anything simple at all; but only for a general dimension of being a right or proper thing to say as opposed to a wrong thing, in these circumstances, to this audience, for these purposes and with these intentions.

In general we may say this: with both statements (and, for example, descriptions) *and* warnings, &c., the question can arise, granting that you had the right to warn and did warn, did state, or did advise, whether you were right to state or warn or advise—not in the sense of whether it was opportune or expedient, but whether, on the facts and your knowledge of the facts and the purposes for which you were speaking, and so on, this was the proper thing to say.

This doctrine is quite different from much that the pragmatists have said, to the effect that the true is what works, &c. The truth or falsity of a statement depends not merely on the meanings of words but on what act you were performing in what circumstances.

What then finally is left of the distinction of the performative and constative utterance? Really we may say that what we had in mind here was this:

(*a*) With the constative utterance, we abstract from the illocutionary (let alone the perlocutionary) aspects of

the speech act, and we concentrate on the locutionary: moreover, we use an over-simplified notion of correspondence with the facts—over-simplified because essentially it brings in the illocutionary aspect. This is the ideal of what would be right to say in all circumstances, for any purpose, to any audience, &c. Perhaps it is sometimes realized.

(*b*) With the performative utterance, we attend as much as possible to the illocutionary force of the utterance, and abstract from the dimension of correspondence with facts.

Perhaps neither of these abstractions is so very expedient: perhaps we have here not really two poles, but rather an historical development. Now in certain cases, perhaps with mathematical formulas in physics books as examples of constatives, or with the issuing of simple executive orders or the giving of simple names, say, as examples of performatives, we approximate in real life to finding such things. It was examples of this kind, like 'I apologize', and 'The cat is on the mat', said for no conceivable reason, extreme marginal cases, that gave rise to the idea of two distinct utterances. But the real conclusion must surely be that we need (*a*) to distinguish between locutionary and illocutionary acts, and (*b*) specially and critically to establish with respect to each kind of illocutionary act—warnings, estimates, verdicts, statements, and descriptions—what if any is the specific way in which they are intended, first to be in order or not in order, and second, to be 'right' or 'wrong'; what terms

of appraisal and disappraisal are used for each and what they mean. This is a wide field and certainly will not lead to a simple distinction of 'true' and 'false'; nor will it lead to a distinction of statements from the rest, for stating is only one among very numerous speech acts of the illocutionary class.

Furthermore, in general the locutionary act as much as the illocutionary is an abstraction only: every genuine speech act is both. (This is similar to the way in which the phatic act, the rhetic act, &c., are mere abstractions.) But, of course, typically we distinguish different abstracted 'acts' by means of the possible slips between cup and lip, that is, in this case, the different types of nonsense which may be engendered in performing them. We may compare with this point what was said in the opening lecture about the classification of kinds of nonsense.

GRICE, H.P.

MEANING

IN: THE PHILOSOPHICAL REVIEW
66 (1970)

MEANING

CONSIDER the following sentences:

"Those spots mean (meant) measles."

"Those spots didn't mean anything to me, but to the doctor they meant measles."

"The recent budget means that we shall have a hard year."

(1) I cannot say, "Those spots meant measles, but he hadn't got measles," and I cannot say, "The recent budget means that we shall have a hard year, but we shan't have." That is to say, in cases like the above, *x meant that p* and *x means that p* entail *p*.

(2) I cannot argue from "Those spots mean (meant) measles" to any conclusion about "what is (was) meant by those spots"; for example, I am not entitled to say, "What was meant by those spots was that he had measles." Equally I cannot draw from the statement about the recent budget the conclusion "What is meant by the recent budget is that we shall have a hard year."

(3) I cannot argue from "Those spots meant measles" to any conclusion to the effect that somebody or other meant by those spots so-and-so. *Mutatis mutandis*, the same is true of the sentence about the recent budget.

(4) For none of the above examples can a restatement be found in which the verb "mean" is followed by a sentence or phrase in inverted commas. Thus "Those spots meant measles" cannot be reformulated as "Those spots meant 'measles' " or as "Those spots meant 'he has measles.' "

(5) On the other hand, for all these examples an approximate restatement can be found beginning with the phrase "The fact that . . ."; for example, "The fact that he had those spots meant that he had measles" and "The fact that the recent budget was as it was means that we shall have a hard year."

Now contrast the above sentences with the following:

"Those three rings on the bell (of the bus) mean that the 'bus is full.' "

"That remark, 'Smith couldn't get on without his trouble and strife,' meant that Smith found his wife indispensable."

(1) I can use the first of these and go on to say, "But it isn't in fact full—the conductor has made a mistake"; and I can use the second and go on, "But in fact Smith deserted her seven years ago." That is to say, here *x means that p* and *x meant that p* do not entail *p*.

(2) I can argue from the first to some statement about "what is (was) meant" by the rings on the bell and from the second to some statement about "what is (was) meant" by the quoted remark.

(3) I can argue from the first sentence to the conclusion that somebody (viz., the conductor) meant, or at any rate should have meant, by the rings that the bus is full, and I can argue analogously for the second sentence.

(4) The first sentence can be restated in a form in which the verb "mean" is followed by a phrase in inverted commas, that is, "Those three rings on the bell mean 'the bus is full.'" So also can the second sentence.

(5) Such a sentence as "The fact that the bell has been rung three times means that the bus is full" is not a restatement of the meaning of the first sentence. Both may be true, but they do not have, even approximately, the same meaning.

When the expressions "means," "means something," "means that" are used in the kind of way in which they are used in the first set of sentences, I shall speak of the sense, or senses, in which they are used, as the *natural* sense, or senses, of the expressions in question. When the expressions are used in the kind of way in which they are used in the second set of sentences, I shall speak of the sense, or senses, in which they are used, as the *nonnatural* sense, or senses, of the expressions in question. I shall use the abbreviation "means$_{NN}$" to distinguish the nonnatural sense or senses.

I propose, for convenience, also to include under the head of natural senses of "mean" such senses of "mean" as may be exemplified in sentences of the pattern "*A* means (meant) *to do* so-and-so (by *x*)," where *A* is a human agent. By contrast, as the previous examples show, I include under the head of non-

natural senses of "mean" any senses of "mean" found in sentences of the patterns "*A* means (meant) something by *x*" or "*A* means (meant) by *x* that. . . ." (This is overrigid; but it will serve as an indication.)

I do not want to maintain that *all* our uses of "mean" fall easily, obviously, and tidily into one of the two groups I have distinguished; but I think that in most cases we should be at least fairly strongly inclined to assimilate a use of "mean" to one group rather than to the other. The question which now arises is this: "What more can be said about the distinction between the cases where we should say that the word is applied in a natural sense and the cases where we should say that the word is applied in an nonnatural sense?" Asking this question will not of course prohibit us from trying to give an explanation of "meaning$_{NN}$" in terms of one or another natural sense of "mean."

This question about the distinction between natural and nonnatural meaning is, I think, what people are getting at when they display an interest in a distinction between "natural" and "conventional" signs. But I think my formulation is better. For some things which can mean$_{NN}$ something are not signs (e.g., words are not), and some are not conventional in any ordinary sense (e.g., certain gestures); while some things which mean naturally are not signs of what they mean (cf. the recent budget example).

I want first to consider briefly, and reject, what I might term a causal type of answer to the question, "What is meaning$_{NN}$?" We might try to say, for instance, more or less with C. L. Stevenson,[1] that for *x* to mean$_{NN}$ something, *x* must have (roughly) a tendency to produce in an audience some attitude (cognitive or otherwise) and a tendency, in the case of a speaker, to *be* produced *by* that attitude, these tendencies being dependent on "an elaborate process of conditioning attending the use of the sign in communication."[2] This clearly will not do.

(1) Let us consider a case where an utterance, if it qualifies at all as meaning$_{NN}$ something, will be of a descriptive or informative kind and the relevant attitude, therefore, will be a cognitive one,

[1] *Ethics and Language* (New Haven, 1944), ch. iii.
[2] *Ibid.*, p. 57.

for example, a belief. (I use "utterance" as a neutral word to apply to any candidate for meaning$_{NN}$; it has a convenient act-object ambiguity.) It is no doubt the case that many people have a tendency to put on a tail coat when they think they are about to go to a dance, and it is no doubt also the case that many people, on seeing someone put on a tail coat, would conclude that the person in question was about to go to a dance. Does this satisfy us that putting on a tail coat means$_{NN}$ that one is about to go to a dance (or indeed means$_{NN}$ anything at all)? Obviously not. It is no help to refer to the qualifying phrase "dependent on an elaborate process of conditioning. . . ." For if all this means is that the response to the sight of a tail coat being put on is in some way learned or acquired, it will not exclude the present case from being one of meaning$_{NN}$. But if we have to take seriously the second part of the qualifying phrase ("attending the use of the sign in communication"), then the account of meaning$_{NN}$ is obviously circular. We might just as well say, "*X* has meaning$_{NN}$ if it is used in communication," which, though true, is not helpful.

(2) If this is not enough, there is a difficulty—really the same difficulty, I think—which Stevenson recognizes: how we are to avoid saying, for example, that "Jones is tall" is part of what is meant by "Jones is an athlete," since to tell someone that Jones is an athlete would tend to make him believe that Jones is tall. Stevenson here resorts to invoking linguistic rules, namely, a permissive rule of language that "athletes may be nontall." This amounts to saying that we are not prohibited by rule from speaking of "nontall athletes." But why are we not prohibited? Not because it is not bad grammar, or is not impolite, and so on, but presumably because it is not meaningless (or, if this is too strong, does not in any way violate the rules of meaning for the expressions concerned). But this seems to involve us in another circle. Moreover, one wants to ask why, if it is legitimate to appeal here to rules to distinguish what is meant from what is suggested, this appeal was not made earlier, in the case of groans, for example, to deal with which Stevenson originally introduced the qualifying phrase about dependence on conditioning.

A further deficiency in a causal theory of the type just

expounded seems to be that, even if we accept it as it stands, we are furnished with an analysis only of statements about the *standard* meaning, or the meaning in general, of a "sign." No provision is made for dealing with statements about what a particular speaker or writer means by a sign on a particular occasion (which may well diverge from the standard meaning of the sign); nor is it obvious how the theory could be adapted to make such provision. One might even go further in criticism and maintain that the causal theory ignores the fact that the meaning (in general) of a sign needs to be explained in terms of what users of the sign do (or should) mean by it on particular occasions; and so the latter notion, which is unexplained by the causal theory, is in fact the fundamental one. I am sympathetic to this more radical criticism, though I am aware that the point is controversial.

I do not propose to consider any further theories of the "causal-tendency" type. I suspect no such theory could avoid difficulties analogous to those I have outlined without utterly losing its claim to rank as a theory of this type.

I will now try a different and, I hope, more promising line. If we can elucidate the meaning of

"*x* meant$_{NN}$ something (on a particular occasion)" and
"*x* meant$_{NN}$ that so-and-so (on a particular occasion)"

and of

"*A* meant$_{NN}$ something by *x* (on a particular occasion)" and
"*A* meant$_{NN}$ by *x* that so-and-so (on a particular occasion),"

this might reasonably be expected to help us with

"*x* means$_{NN}$ (timeless) something (that so-and-so),"
"*A* means$_{NN}$ (timeless) by *x* something (that so-and-so),"

and with the explication of "means the same as," "understands," "entails," and so on. Let us for the moment pretend that we have to deal only with utterances which might be informative or descriptive.

A first shot would be to suggest that "*x* meant$_{NN}$ something" would be true if *x* was intended by its utterer to induce a belief in some "audience" and that to say what the belief was would be to say what *x* meant$_{NN}$. This will not do. I might leave *B*'s

handkerchief near the scene of a murder in order to induce the detective to believe that B was the murderer; but we should not want to say that the handkerchief (or my leaving it there) $meant_{NN}$ anything or that I had $meant_{NN}$ by leaving it that B was the murderer. Clearly we must at least add that, for x to have $meant_{NN}$ anything, not merely must it have been "uttered" with the intention of inducing a certain belief but also the utterer must have intended an "audience" to recognize the intention behind the utterance.

This, though perhaps better, is not good enough. Consider the following cases:

(1) Herod presents Salome with the head of St. John the Baptist on a charger.

(2) Feeling faint, a child lets its mother see how pale it is (hoping that she may draw her own conclusions and help).

(3) I leave the china my daughter has broken lying around for my wife to see.

Here we seem to have cases which satisfy the conditions so far given for $meaning_{NN}$. For example, Herod intended to make Salome believe that St. John the Baptist was dead and no doubt also intended Salome to recognize that he intended her to believe that St. John the Baptist was dead. Similarly for the other cases. Yet I certainly do not think that we should want to say that we have here cases of $meaning_{NN}$.

What we want to find is the difference between, for example, "deliberately and openly letting someone know" and "telling" and between "getting someone to think" and "telling."

The way out is perhaps as follows. Compare the following two cases:

(1) I show Mr. X a photograph of Mr. Y displaying undue familiarity to Mrs. X.

(2) I draw a picture of Mr. Y behaving in this manner and show it to Mr. X.

I find that I want to deny that in (1) the photograph (or my showing it to Mr. X) $meant_{NN}$ anything at all; while I want to assert that in (2) the picture (or my drawing and showing it)

$meant_{NN}$ something (that Mr. Y had been unduly unfamiliar), or at least that I had $meant_{NN}$ by it that Mr. Y had been unduly familiar. What is the difference between the two cases? Surely that in case (1) Mr. X's recognition of my intention to make him believe that there is something between Mr. Y and Mrs. X is (more or less) irrelevant to the production of this effect by the photograph. Mr. X would be led by the photograph at least to suspect Mrs. X even if instead of showing it to him I had left it in his room by accident; and I (the photograph shower) would not be unaware of this. But it will make a difference to the effect of my picture on Mr. X whether or not he takes me to be intending to inform him (make him believe something) about Mrs. X, and not to be just doodling or trying to produce a work of art.

But now we seem to be landed in a further difficulty if we accept this account. For consider now, say, frowning. If I frown spontaneously, in the ordinary course of events, someone looking at me may well treat the frown as a natural sign of displeasure. But if I frown deliberately (to convey my displeasure), an onlooker may be expected, provided he recognizes my intention, *still* to conclude that I am displeased. Ought we not then to say, since it could not be expected to make any difference to the onlooker's reaction whether he regards my frown as spontaneous or as intended to be informative, that my frown (deliberate) does *not* $mean_{NN}$ anything? I think this difficulty can be met; for though in general a deliberate frown may have the same effect (as regards inducing belief in my displeasure) as a spontaneous frown, it can be expected to have the same effect only *provided* the audience takes it as intended to convey displeasure. That is, if we take away the recognition of intention, leaving the other circumstances (including the recognition of the frown as deliberate), the belief-producing tendency of the frown must be regarded as being impaired or destroyed.

Perhaps we may sum up what is necessary for A to mean something by x as follows. A must intend to induce by x a belief in an audience, and he must also intend his utterance to be recognized as so intended. But these intentions are not independent; the recognition is intended by A to play its part in inducing the belief, and if it does not do so something will have gone wrong

with the fulfillment of *A*'s intentions. Moreover, *A*'s intending that the recognition should play this part implies, I think, that he assumes that there is some chance that it will in fact play this part, that he does not regard it as a foregone conclusion that the belief will be induced in the audience whether or not the intention behind the utterance is recognized. Shortly, perhaps, we may say that "*A* meant$_{NN}$ something by *x*" is roughly equivalent to "*A* uttered *x* with the intention of inducing a belief by means of the recognition of this intention." (This seems to involve a reflexive paradox, but it does not really do so.)

Now perhaps it is time to drop the pretense that we have to deal only with "informative" cases. Let us start with some examples of imperatives or quasi-imperatives. I have a very avaricious man in my room, and I want him to go; so I throw a pound note out of the window. Is there here any utterance with a meaning$_{NN}$? No, because in behaving as I did, I did not intend his recognition of my purpose to be in any way effective in getting him to go. This is parallel to the photograph case. If on the other hand I had pointed to the door or given him a little push, then my behavior might well be held to constitute a meaningful$_{NN}$ utterance, just because the recognition of my intention would be intended by me to be effective in speeding his departure. Another pair of cases would be (1) a policeman who stops a car by standing in its way and (2) a policeman who stops a car by waving.

Or, to turn briefly to another type of case, if as an examiner I fail a man, I may well cause him distress or indignation or humiliation; and if I am vindictive, I may intend this effect and even intend him to recognize my intention. But I should not be inclined to say that my failing him meant$_{NN}$ anything. On the other hand, if I cut someone in the street I do feel inclined to assimilate this to the cases of meaning$_{NN}$, and this inclination seems to me dependent on the fact that I could not reasonably expect him to be distressed (indignant, humiliated) unless he recognized my intention to affect him in this way. (Cf., if my college stopped my salary altogether I should accuse them of ruining me; if they cut it by 2/6^{d} I might accuse them of insulting me; with some intermediate amounts I might not know quite what to say.)

Perhaps then we may make the following generalizations.

(1) "*A* meant$_{NN}$ something by *x*" is (roughly) equivalent to "*A* intended the utterance of *x* to produce some effect in an audience by means of the recognition of this intention"; and we may add that to ask what *A* meant is to ask for a specification of the intended effect (though, of course, it may not always be possible to get a straight answer involving a "that" clause, for example, "a belief that . . .").

(2) "*x* meant something" is (roughly) equivalent to "Somebody meant$_{NN}$ something by *x*." Here again there will be cases where this will not quite work. I feel inclined to say that (as regards traffic lights) the change to red meant$_{NN}$ that the traffic was to stop; but it would be very unnatural to say, "Somebody (e.g., the Corporation) meant$_{NN}$ by the red-light change that the traffic was to stop." Nevertheless, there seems to be *some* sort of reference to somebody's intentions.

(3) "*x* means$_{NN}$ (timeless) that so-and-so" might as a first shot be equated with some statement or disjunction of statements about what "people" (vague) intend (with qualifications about "recognition") to effect by *x*. I shall have a word to say about this.

Will any kind of intended effect do, or may there be cases where an effect is intended (with the required qualifications) and yet we should not want to talk of meaning$_{NN}$? Suppose I discovered some person so constituted that, when I told him that whenever I grunted in a special way I wanted him to blush or to incur some physical malady, thereafter whenever he recognized the grunt (and with it my intention), he did blush or incur the malady. Should we then want to say that the grunt meant$_{NN}$ something? I do not think so. This points to the fact that for *x* to have meaning$_{NN}$, the intended effect must be something which in some sense is within the control of the audience, or that in some sense of "reason" the recognition of the intention behind *x* is for the audience a reason and not merely a cause. It might look as if there is a sort of pun here ("reason for believing" and "reason for doing"), but I do not think this is serious. For though no doubt from one point of view questions about reasons for believing are questions about evidence and so quite different from questions

about reasons for doing, nevertheless to recognize an utterer's intention in uttering x (descriptive utterance), to have a reason for believing that so-and-so, is at least quite like "having a motive for" accepting so-and-so. Decisions "that" seem to involve decisions "to" (and this is why we can "refuse to believe" and also be "compelled to believe"). (The "cutting" case needs slightly different treatment, for one cannot in any straightforward sense "decide" to be offended; but one can refuse to be offended.) It looks then as if the intended effect must be something within the control of the audience, or at least the *sort* of thing which is within its control.

One point before passing to an objection or two. I think it follows that from what I have said about the connection between meaning$_{NN}$ and recognition of intention that (insofar as I am right) only what I may call the primary intention of an utterer is relevant to the meaning$_{NN}$ of an utterance. For if I utter x, intending (with the aid of the recognition of this intention) to induce an effect E, and intend this effect E to lead to a further effect F, then insofar as the occurrence of F is thought to be dependent solely on E, I cannot regard F as in the least dependent on recognition of my intention to induce E. That is, if (say) I intend to get a man to do something by giving him some information, it cannot be regarded as relevant to the meaning$_{NN}$ of my utterance to describe what I intend him to do.

Now some question may be raised about my use, fairly free, of such words as "intention" and "recognition." I must disclaim any intention of peopling all our talking life with armies of complicated psychological occurrences. I do not hope to solve any philosophical puzzles about intending, but I do want briefly to argue that no special difficulties are raised by my use of the word "intention" in connection with meaning. First, there will be cases where an utterance is accompanied or preceded by a conscious "plan," or explicit formulation of intention (e.g., I declare how I am going to use x, or ask myself how to "get something across"). The presence of such an explicit "plan" obviously counts fairly heavily in favor of the utterer's intention (meaning) being as "planned"; though it is not, I think, conclusive; for example, a speaker who has declared an intention

to use a familiar expression in an unfamiliar way may slip into the familiar use. Similarly in nonlinguistic cases: if we are asking about an agent's intention, a previous expression counts heavily; nevertheless, a man might plan to throw a letter in the dustbin and yet take it to the post; when lifting his hand he might "come to" and say *either* "I didn't intend to do this at all" *or* "I suppose I must have been intending to put it in."

Explicitly formulated linguistic (or quasi-linguistic) intentions are no doubt comparatively rare. In their absence we would seem to rely on very much the same kinds of criteria as we do in the case of nonlinguistic intentions where there is a general usage. An utterer is held to intend to convey what is normally conveyed (or normally intended to be conveyed), and we require a good reason for accepting that a particular use diverges from the general usage (e.g., he never knew or had forgotten the general usage). Similarly in nonlinguistic cases: we are presumed to intend the normal consequences of our actions.

Again, in cases where there is doubt, say, about which of two or more things an utterer intends to convey, we tend to refer to the context (linguistic or otherwise) of the utterance and ask which of the alternatives would be relevant to other things he is saying or doing, or which intention in a particular situation would fit in with some purpose he obviously has (e.g., a man who calls for a "pump" at a fire would not want a bicycle pump). Nonlinguistic parallels are obvious: context is a criterion in settling the question of why a man who has just put a cigarette in his mouth has put his hand in his pocket; relevance to an obvious end is a criterion in settling why a man is running away from a bull.

In certain linguistic cases we ask the utterer afterward about his intention, and in a few of these cases (the very difficult ones, like a philosopher asked to explain the meaning of an unclear passage in one of his works), the answer is not based on what he remembers but is more like a decision, a decision about how what he said is to be taken. I cannot find a nonlinguistic parallel here; but the case is so special as not to seem to contribute a vital difference.

All this is very obvious; but surely to show that the criteria

for judging linguistic intentions are very like the criteria for judging nonlinguistic intentions is to show that linguistic intentions are very like nonlinguistic intentions.

H. P. Grice

St. John's College
Oxford

KRIPKE, SAUL

WITTGENSTEIN ON RULES AND PRIVATE LANGUAGE

I

Introductory

Wittgenstein's celebrated argument against 'private language' has been discussed so often that the utility of yet another exposition is certainly open to question. Most of the exposition which follows occurred to the present writer some time ago, in the academic year 1962–3. At that time this approach to Wittgenstein's views struck the present writer with the force of a revelation: what had previously seemed to me to be a somewhat loose argument for a fundamentally implausible conclusion based on dubious and controversial premises now appeared to me to be a powerful argument, even if the conclusions seemed even more radical and, in a sense, more implausible, than before. I thought at that time that I had seen Wittgenstein's argument from an angle and emphasis very different from the approach which dominated standard expositions. Over the years I came to have doubts. First of all, at times I became unsure that I could formulate Wittgenstein's elusive position as a clear argument. Second, the elusive nature of the subject made it possible to interpret some of the standard literature as perhaps seeing the argument in the same way after all. More important, conversations over the years showed that, increasingly, others were seeing the argument with the emphases I preferred. Nevertheless, recent expositions by very able interpreters differ enough from the

following to make me think that a new exposition may still be of use.[1]

A common view of the 'private language argument' in *Philosophical Investigations* assumes that it begins with section 243, and that it continues in the sections immediately following.[2] This view takes the argument to deal primarily with a problem about 'sensation language'. Further discussion of the argument in this tradition, both in support and in criticism, emphasizes such questions as whether the argument invokes a form of the verification principle, whether the form in question is justified, whether it is applied correctly to sensation language, whether the argument rests on an exaggerated scepticism about memory, and so on. Some

[1] Looking through some of the most distinguished commentaries on Wittgenstein of the last ten or fifteen years, I find some that still treat the discussion of rules cursorily, virtually not at all, as if it were a minor topic. Others, who discuss both Wittgenstein's views on the philosophy of mathematics and his views on sensations in detail, treat the discussion of rules as if it were important for Wittgenstein's views on mathematics and logical necessity but separate it from 'the private language argument'. Since Wittgenstein has more than *one* way of arguing for a given conclusion, and even of presenting a single argument, to defend the present exegesis I need not necessarily argue that these other commentaries are in error. Indeed, they may give important and illuminating expositions of facets of the *Investigations* and its argument deemphasized or omitted in this essay. Nevertheless, in emphasis they certainly differ considerably from the present exposition.

[2] Unless otherwise specified (explicitly or contextually), references are to *Philosophical Investigations*. The small numbered units of the *Investigations* are termed 'sections' (or 'paragraphs'). Page references are used only if a section reference is not possible, as in the second part of the *Investigations*. Throughout I quote the standard printed English translation (by G. E. M. Anscombe) and make no attempt to question it except in a very few instances. *Philosophical Investigations* (x+232 pp., parallel German and English text) has undergone several editions since its first publication in 1953 but the paragraphing and pagination remain the same. The publishers are Basil Blackwell, Oxford and Macmillan, New York.

This essay does not proceed by giving detailed exegesis of Wittgenstein's text but rather develops the arguments in its own way. I recommended that the reader reread the *Investigations* in the light of the present exegesis and see whether it illuminates the text.

crucial passages in the discussion following §243 – for example, such celebrated sections as §258 and §265 – have been notoriously obscure to commentators, and it has been thought that their proper interpretation would provide the key to the 'private language argument'.

In my view, the real 'private language argument' is to be found in the sections *preceding* §243. Indeed, in §202 *the conclusion is already stated explicitly*: "Hence it is not possible to obey a rule 'privately': otherwise thinking one was obeying a rule would be the same thing as obeying it." I do not think that Wittgenstein here thought of himself as *anticipating* an argument he was to give in greater detail later. On the contrary, the crucial considerations are all contained in the discussion leading up to the conclusion stated in §202. The sections following §243 are meant to be read in the light of the preceding discussion; difficult as they are in any case, they are much less likely to be understood if they are read in isolation. The 'private language argument' as applied to *sensations* is only a special case of much more general considerations about language previously argued; sensations have a crucial role as an (apparently) convincing *counterexample* to the general considerations previously stated. Wittgenstein therefore goes over the ground again in this special case, marshalling new specific considerations appropriate to it. It should be borne in mind that *Philosophical Investigations* is not a systematic philosophical work where conclusions, once definitely established, need not be reargued. Rather the *Investigations* is written as a perpetual dialectic, where persisting worries, expressed by the voice of the imaginary interlocutor, are never definitively silenced. Since the work is not presented in the form of a deductive argument with definitive theses as conclusions, the same ground is covered repeatedly, from the point of view of various special cases and from different angles, with the hope that the entire process will help the reader see the problems rightly.

The basic structure of Wittgenstein's approach can be presented briefly as follows: A certain problem, or in Humean

terminology, a 'sceptical paradox', is presented concerning the notion of a rule. Following this, what Hume would have called a 'sceptical solution' to the problem is presented. There are two areas in which the force, both of the paradox and of its solution, are most likely to be ignored, and with respect to which Wittgenstein's basic approach is most likely to seem incredible. One such area is the notion of a mathematical rule, such as the rule for addition. The other is our talk of our own inner experience, of sensations and other inner states. In treating both these cases, we should bear in mind the basic considerations about rules and language. Although Wittgenstein has already discussed these basic considerations in considerable generality, the structure of Wittgenstein's work is such that the special cases of mathematics and psychology are not simply discussed by citing a general 'result' already established, but by going over these special cases in detail, in the light of the previous treatment of the general case. By such a discussion, it is hoped that both mathematics and the mind can be seen rightly: since the temptations to see them wrongly arise from the neglect of the same basic considerations about rules and language, the problems which arise can be expected to be analogous in the two cases. In my opinion, Wittgenstein did not view his dual interests in the philosophy of mind and the philosophy of mathematics as interests in two separate, at best loosely related, subjects, as someone might be interested both in music and in economics. Wittgenstein thinks of the two subjects as involving the same basic considerations. For this reason, he calls his investigation of the foundations of mathematics "analogous to our investigation of psychology" (p. 232). It is no accident that essentially the same basic material on rules is included in both *Philosophical Investigations* and in *Remarks on the Foundations of Mathematics*,[3] both times as

[3] Basil Blackwell, Oxford, 1956, xix+204 pp. In the first edition of *Remarks on the Foundations of Mathematics* the editors assert (p. vi) that Wittgenstein appears originally to have intended to include some of the material on mathematics in *Philosophical Investigations*.

The third edition (1978) includes more material than earlier editions,

the basis of the discussions of the philosophies of mind and of mathematics, respectively, which follow.

In the following, I am largely trying to present Wittgenstein's argument, or, more accurately, that set of problems and arguments which I personally have gotten out of reading Wittgenstein. With few exceptions, I am *not* trying to present views of my own; neither am I trying to endorse or to criticize Wittgenstein's approach. In some cases, I have found a precise statement of the problems and conclusions to be elusive. Although one has a strong sense that there is a problem, a rigorous statement of it is difficult. I am inclined to think that Wittgenstein's later philosophical style, and the difficulty he found (see his Preface) in welding his thought into a conventional work presented with organized arguments and conclusions, is not simply a stylistic and literary preference, coupled with a *penchant* for a certain degree of obscurity,[4] but stems in part from the nature of his subject.[5]

I suspect – for reasons that will become clearer later – that to attempt to present Wittgenstein's argument precisely is to some extent to falsify it. Probably many of my formulations and recastings of the argument are done in a way Wittgenstein would not himself approve.[6] So the present paper should be thought of as expounding neither 'Wittgenstein's' argument nor 'Kripke's': rather Wittgenstein's argument as it struck Kripke, as it presented a problem for him.

As I have said, I think the basic 'private language argument' *precedes* section 243, though the sections following 243 are no

and rearranges some of the sections and divisions of earlier editions. When I wrote the present work, I used the first edition. Where the references differ, the equivalent third edition reference is given in square brackets.

[4] Personally I feel, however, that the role of stylistic considerations here cannot be denied. It is clear that purely stylistic and literary considerations meant a great deal to Wittgenstein. His own stylistic preference obviously contributes to the difficulty of his work as well as to its beauty.

[5] See the discussion of this point on page 69 below.

[6] See again the same discussion on page 69.

doubt of fundamental importance as well. I propose to discuss the problem of 'private language' initially without mentioning these latter sections *at all*. Since these sections are often thought to *be* the 'private language argument', to some such a procedure may seem to be a presentation of Hamlet without the prince. Even if this is so, there are many other interesting characters to play.[7]

[7] Looking over what I have written below, I find myself worried that the reader may lose the main thread of Wittgenstein's argument in the extensive treatment of finer points. In particular, the treatment of the dispositional theory below became so extensive because I heard it urged more than once as an answer to the sceptical paradox. That discussion may contain somewhat more of Kripke's argumentation in support of Wittgenstein rather than exposition of Wittgenstein's own argument than does most of the rest of this essay. (See notes 19 and 24 for *some* of the connections. The argument is, however, inspired by Wittgenstein's original text. Probably the argument below with the least direct inspiration from Wittgenstein's text is the argument that our dispositions, like our actual performance, are not potentially infinite. Even this, however, obviously has its origin in Wittgenstein's parallel emphasis on the fact that we explicitly think of only finitely many cases of any rule.) I urge the reader to concentrate, on a first reading, on understanding the intuitive force of Wittgenstein's sceptical problem and to regard byways such as these as secondary.

2

The Wittgensteinian Paradox

In §201, Wittgenstein says, "this was our paradox: no course of action could be determined by a rule, because every course of action can be made to accord with the rule." In this section of the present essay, in my own way I will attempt to develop the 'paradox' in question. The 'paradox' is perhaps the central problem of *Philosophical Investigations*. Even someone who disputes the conclusions regarding 'private language', and the philosophies of mind, mathematics, and logic, that Wittgenstein draws from his problem, might well regard the problem itself as an important contribution to philosophy. It may be regarded as a new form of philosophical scepticism.

Following Wittgenstein, I will develop the problem initially with respect to a mathematical example, though the relevant sceptical problem applies to all meaningful uses of language. I, like almost all English speakers, use the word 'plus' and the symbol '+' to denote a well-known mathematical function, addition. The function is defined for all pairs of positive integers. By means of my external symbolic representation and my internal mental representation, I 'grasp' the rule for addition. One point is crucial to my 'grasp' of this rule. Although I myself have computed only finitely many sums in the past, the rule determines my answer for indefinitely many new sums that I have never previously considered. This is the

whole point of the notion that in learning to add I grasp a rule: my past intentions regarding addition determine a unique answer for indefinitely many new cases in the future.

Let me suppose, for example, that '68 + 57' is a computation that I have never performed before. Since I have performed – even silently to myself, let alone in my publicly observable behavior – only finitely many computations in the past, such an example surely exists. In fact, the same finitude guarantees that there is an example exceeding, in both its arguments, all previous computations. I shall assume in what follows that '68 + 57' serves for this purpose as well.

I perform the computation, obtaining, of course, the answer '125'. I am confident, perhaps after checking my work, that '125' is the correct answer. It is correct both in the arithmetical sense that 125 is the sum of 68 and 57, and in the metalinguistic sense that 'plus', as I intended to use that word in the past, denoted a function which, when applied to the numbers I called '68' and '57', yields the value 125.

Now suppose I encounter a bizarre sceptic. This sceptic questions my certainty about my answer, in what I just called the 'metalinguistic' sense. Perhaps, he suggests, as I used the term 'plus' in the past, the answer I intended for '68+57' should have been '5'! Of course the sceptic's suggestion is obviously insane. My initial response to such a suggestion might be that the challenger should go back to school and learn to add. Let the challenger, however, continue. After all, he says, if I am now so confident that, as I used the symbol '+', my intention was that '68+57' should turn out to denote 125, this cannot be because I explicitly gave myself instructions that 125 is the result of performing the addition in this particular instance. By hypothesis, I did no such thing. But of course the idea is that, in this new instance, I should apply the very same function or rule that I applied so many times in the past. But who is to say what function this was? In the past I gave myself only a finite number of examples instantiating this function. All, we have supposed, involved numbers smaller than 57. So perhaps in the past I used 'plus' and '+' to denote a function

which I will call 'quus' and symbolize by '⊕'. It is defined by:

$$x \oplus y = x + y, \text{ if } x, y < 57$$
$$= 5 \quad \text{otherwise.}$$

Who is to say that this is not the function I previously meant by '+'?

The sceptic claims (or feigns to claim) that I am now misinterpreting my own previous usage. By 'plus', he says, I *always meant* quus;[8] now, under the influence of some insane frenzy, or a bout of LSD, I have come to misinterpret my own previous usage.

Ridiculous and fantastic though it is, the sceptic's hypothesis is not logically impossible. To see this, assume the common sense hypothesis that by '+' I *did* mean addition. Then it would be *possible*, though surprising, that under the influence of a momentary 'high', I should misinterpret all my past uses of the plus sign as symbolizing the quus function, and proceed, in conflict with my previous linguistic intentions, to compute 68 plus 57 as 5. (I would have made a mistake, not in mathematics, but in the supposition that I had accorded with my previous linguistic intentions.) The sceptic is proposing that I have made a mistake precisely of this kind, but with a plus and quus reversed.

Now if the sceptic proposes his hypothesis sincerely, he is crazy; such a bizarre hypothesis as the proposal that I always meant quus is absolutely wild. Wild it indubitably is, no doubt it is false; but if it is false, there must be some fact about my past usage that can be cited to refute it. For although the hypothesis is wild, it does not seem to be *a priori* impossible.

[8] Perhaps I should make a remark about such expressions as "By 'plus' I meant quus (or plus)," "By 'green' I meant green," etc. I am not familiar with an accepted felicitous convention to indicate the object of the verb 'to mean'. There are two problems. First, if one says, "By 'the woman who discovered radium' I meant the woman who discovered radium," the object can be interpreted in two ways. It may stand for a woman (Marie Curie), in which case the assertion is true only if 'meant' is used to mean referred to (as it can be used); or it may be used to denote the *meaning* of the quoted expression, not a woman, in which case the assertion is true

Of course this bizarre hypothesis, and the references to LSD, or to an insane frenzy, are in a sense merely a dramatic device. The basic point is this. Ordinarily, I suppose that, in computing '68+57' as I do, I do not simply make an unjustified leap in the dark. I follow directions I previously gave myself that uniquely determine that in this new instance I should say '125'. What are these directions? By hypothesis, I never explicitly told myself that I should say '125' in this very instance. Nor can I say that I should simply 'do the same thing

with 'meant' used in the ordinary sense. Second, as is illustrated by 'referred to', 'green', 'quus', etc. above, as objects of 'meant', one must use various expressions as objects in an awkward manner contrary to normal grammar. (Frege's difficulties concerning unsaturatedness are related.) Both problems tempt one to put the object in quotation marks, like the subject; but such a usage conflicts with the convention of philosophical logic that a quotation denotes the expression quoted. Some special 'meaning marks', as proposed for example by David Kaplan, could be useful here. If one is content to ignore the first difficulty and always use 'mean' to mean denote (for most purposes of the present paper, such a reading would suit at least as well as an intensional one; often I speak as if it is a *numerical function* that is meant by plus), the second problem might lead one to nominalize the objects – 'plus' denotes the plus function, 'green' denotes greenness, etc. I contemplated using italics ("'plus' means *plus*"; "'mean' may mean *denote*"), but I decided that normally (except when italics are otherwise appropriate, especially when a neologism like 'quus' is introduced for the first time), I will write the object of 'to mean' as an ordinary roman object. The convention I have adopted reads awkwardly in the written language but sounds rather reasonable in the spoken language.

Since use–mention distinctions are significant for the argument as I give it, I try to remember to use quotation marks when an expression is mentioned. However, quotation marks are also used for other purposes where they might be invoked in normal non-philosophical English writing (for example, in the case of "'meaning marks'" in the previous paragraph, or "'quasi-quotation'" in the next sentence). Readers familiar with Quine's 'quasi-quotation' will be aware that in some cases I use ordinary quotation where logical purity would require that I use quasi-quotation or some similar device. I have not tried to be careful about this matter, since I am confident that in practice readers will not be confused.

I always did,' if this means 'compute according to the rule exhibited by my previous examples.' That rule could just as well have been the rule for quaddition (the quus function) as for addition. The idea that in fact quaddition *is* what I meant, that in a sudden frenzy I have changed my previous usage, dramatizes the problem.

In the discussion below the challenge posed by the sceptic takes two forms. First, he questions whether there is any *fact* that I meant plus, not quus, that will answer his sceptical challenge. Second, he questions whether I have any reason to be so confident that now I should answer '125' rather than '5'. The two forms of the challenge are related. I am confident that I should answer '125' because I am confident that this answer also accords with what I *meant*. Neither the accuracy of my computation nor of my memory is under dispute. So it ought to be agreed that *if* I meant plus, then unless I wish to change my usage, I am justified in answering (indeed compelled to answer) '125', not '5'. An answer to the sceptic must satisfy two conditions. First, it must give an account of what fact it is (about my mental state) that constitutes my meaning plus, not quus. But further, there is a condition that any putative candidate for such a fact must satisfy. It must, in some sense, show how I am justified in giving the answer '125' to '68+57'. The 'directions' mentioned in the previous paragraph, that determine what I should do in each instance, must somehow be 'contained' in any candidate for the fact as to what I meant. Otherwise, the sceptic has not been answered when he holds that my present response is arbitrary. Exactly how this condition operates will become much clearer below, after we discuss Wittgenstein's paradox on an intuitive level, when we consider various philosophical theories as to what the fact that I meant plus might consist in. There will be many specific objections to these theories. But all fail to give a candidate for a fact as to what I meant that would show that only '125', not '5', is the answer I 'ought' to give.

The ground rules of our formulation of the problem should be made clear. For the sceptic to converse with me at all, we

must have a common language. So I am supposing that the sceptic, provisionally, is not questioning my *present* use of the word 'plus'; he agrees that, according to my *present* usage, '68 plus 57' denotes 125. Not only does he agree with me on this, he conducts the entire debate with me in my language as I *presently* use it. He merely questions whether my present usage agrees with my past usage, whether I am *presently* conforming to my *previous* linguistic intentions. The problem is not "How do I know that 68 plus 57 is 125?", which should be answered by giving an arithmetical computation, but rather "How do I know that '68 plus 57', as I *meant* 'plus' in the *past*, should denote 125?" If the word 'plus' as I used it in the past, denoted the quus function, not the plus function ('quaddition' rather than addition), then my *past* intention was such that, asked for the value of '68 plus 57', I should have replied '5'.

I put the problem in this way so as to avoid confusing questions about whether the discussion is taking place 'both inside and outside language' in some illegitimate sense.[9] If we are querying the meaning of the word 'plus', how can we use it (and variants, like 'quus') at the same time? So I suppose that the sceptic assumes that he and I agree in our *present* uses of the word 'plus': we both use it to denote addition. He does *not* – at least initially – deny or doubt that addition is a genuine function, defined on all pairs of integers, nor does he deny that we can speak of it. Rather he asks why I now believe that by 'plus' in the *past*, I meant addition rather than quaddition. If I meant the former, then to accord with my previous usage I should say '125' when asked to give the result of calculating '68 plus 57'. If I meant the latter, I should say '5'.

The present exposition tends to differ from Wittgenstein's original formulations in taking somewhat greater care to make explicit a distinction between use and mention, and between questions about present and past usage. About the present example Wittgenstein might simply ask, "How do I know that I should respond '125' to the query '68+57'?" or "How do

[9] I believe I got the phrase "both inside and outside language" from a conversation with Rogers Albritton.

I know that '68+57' comes out 125?" I have found that when the problem is formulated this way, some listeners hear it as a sceptical problem about *arithmetic*: "How do I know that 68+57 is 125?" (Why not answer this question with a mathematical proof?) At least at this stage, scepticism about arithmetic should not be taken to be in question: we may assume, if we wish, that 68+57 *is* 125. Even if the question is reformulated 'metalinguistically' as "How do I know that 'plus', as I use it, denotes a function that, when applied to 68 and 57, yields 125?", one may answer, "Surely I know that 'plus' denotes the plus function and accordingly that '68 plus 57' denotes 68 plus 57. But if I know arithmetic, I know that 68 plus 57 is 125. So I know that '68 plus 57' denotes 125!" And surely, if I use language at all, I cannot doubt coherently that 'plus', as I now use it, denotes plus! Perhaps I cannot (at least at this stage) doubt this about my *present* usage. But I can doubt that my *past* usage of 'plus' denoted plus. The previous remarks – about a frenzy and LSD – should make this quite clear.

Let me repeat the problem. The sceptic doubts whether any instructions I gave myself in the past compel (or justify) the answer '125' rather than '5'. He puts the challenge in terms of a sceptical hypothesis about a change in my usage. Perhaps when I used the term 'plus' in the *past*, I always meant quus: by hypothesis I never gave myself any explicit directions that were incompatible with such a supposition.

Of course, ultimately, if the sceptic is right, the concepts of meaning and of intending one function rather than another will make no sense. For the sceptic holds that no fact about my past history – nothing that was ever in my mind, or in my external behavior – establishes that I meant plus rather than quus. (Nor, of course, does any fact establish that I meant quus!) But if this is correct, there can of course be no fact about which function I meant, and if there can be no fact about which particular function I meant in the *past*, there can be none in the *present* either. But before we pull the rug out from under our own feet, we begin by speaking as if the notion that at present

we mean a certain function by 'plus' is unquestioned and unquestionable. Only *past* usages are to be questioned. Otherwise, we will be unable to *formulate* our problem.

Another important rule of the game is that there are no limitations, in particular, no *behaviorist* limitations, on the facts that may be cited to answer the sceptic. The evidence is not to be confined to that available to an external observer, who can observe my overt behavior but not my internal mental state. It would be interesting if nothing in my external behavior could show whether I meant plus or quus, but something about my inner state could. But the problem here is more radical. Wittgenstein's philosophy of mind has often been viewed as behavioristic, but to the extent that Wittgenstein may (or may not) be hostile to the 'inner', no such hostility is to be assumed as a premise; it is to be argued as a conclusion. So whatever 'looking into my mind' may be, the sceptic asserts that even if God were to do it, he still could not determine that I meant addition by 'plus'.

This feature of Wittgenstein contrasts, for example, with Quine's discussion of the 'indeterminacy of translation'.[10] There are many points of contact between Quine's discussion and Wittgenstein's. Quine, however, is more than content to assume that only behavioral evidence is to be admitted into his discussion. Wittgenstein, by contrast, undertakes an extensive introspective[11] investigation, and the results of the investiga-

[10] See W. V. Quine, *Word and Object* (MIT, The Technology Press, Cambridge, Massachusetts, 1960, xi+294 pp.), especially chapter 2, 'Translation and Meaning' (pp. 26–79). See also *Ontological Relativity and Other Essays* (Columbia University Press, New York and London, 1969, viii+165 pp.), especially the first three chapters (pp. 1–90); and see also "On the Reasons for the Indeterminacy of Translation," *The Journal of Philosophy*, vol. 67 (1970), pp. 178–83.

Quine's views are discussed further below, see pp. 55–7.

[11] I do not mean the term 'introspective' to be laden with philosophical doctrine. Of course much of the baggage that has accompanied this term would be objectionable to Wittgenstein in particular. I simply mean that he makes use, in his discussion, of our own memories and knowledge of our 'inner' experiences.

tion, as we shall see, form a key feature of his argument. Further, the way the sceptical doubt is presented is not behavioristic. It is presented from the 'inside'. Whereas Quine presents the problem about meaning in terms of a linguist, trying to guess what someone *else* means by his words on the basis of his behavior, Wittgenstein's challenge can be presented to me as a question about *myself*: was there some past fact about me – what I 'meant' by plus – that mandates what I do now?

To return to the sceptic. The sceptic argues that when I answered '125' to the problem '68+57', my answer was an unjustified leap in the dark; my past mental history is equally compatible with the hypothesis that I meant quus, and therefore should have said '5'. We can put the problem this way: When asked for the answer to '68+57', I unhesitatingly and automatically produced '125', but it would seem that if previously I never performed this computation explicitly I might just as well have answered '5'. Nothing justifies a brute inclination to answer one way rather than another.

Many readers, I should suppose, have long been impatient to protest that our problem arises only because of a ridiculous model of the instruction I gave myself regarding 'addition'. Surely I did not merely give myself some finite number of examples, from which I am supposed to extrapolate the whole table ("Let '+' be the function instantiated by the following examples: . . ."). No doubt infinitely many functions are compatible with *that*. Rather I learned – and internalized instructions for – a *rule* which determines how addition is to be continued. What was the rule? Well, say, to take it in its most primitive form: suppose we wish to add x and y. Take a huge bunch of marbles. First count out x marbles in one heap. Then count out y marbles in another. Put the two heaps together and count out the number of marbles in the union thus formed. The result is $x+y$. This set of directions, I may suppose, I explicitly gave myself at some earlier time. It is engraved on my mind as on a slate. It is incompatible with the hypothesis that I meant quus. It is this set of directions, not the finite list of

particular additions I performed in the past, that justifies and determines my present response. This consideration is, after all, reinforced when we think what I really *do* when I add 68 and 57. I do not reply automatically with the answer '125' nor do I consult some non-existent past instructions that I should answer '125' in this case. Rather I proceed according to an *algorithm* for addition that I previously learned. The algorithm is more sophisticated and practically applicable than the primitive one just described, but there is no difference in principle.

Despite the initial plausibility of this objection, the sceptic's response is all too obvious. True, if 'count', as I used the word in the past, referred to the act of counting (and my other past words are correctly interpreted in the standard way), then 'plus' must have stood for addition. But I applied 'count', like 'plus', to only finitely many past cases. Thus the sceptic can question my present interpretation of my past usage of 'count' as he did with 'plus'. In particular, he can claim that by 'count' I formerly meant *quount*, where to 'quount' a heap is to count it in the ordinary sense, unless the heap was formed as the union of two heaps, one of which has 57 or more items, in which case one must automatically give the answer '5'. It is clear that if in the past 'counting' meant quounting, and if I follow the rule for 'plus' that was quoted so triumphantly to the sceptic, I must admit that '68+57' must yield the answer '5'. Here I have supposed that previously 'count' was never applied to heaps formed as the union of sub-heaps either of which has 57 or more elements, but if this particular upper bound does not work, another will do. For the point is perfectly general: if 'plus' is explained in terms of 'counting', a non-standard interpretation of the latter will yield a non-standard interpretation of the former.[12]

[12] The same objection scotches a related suggestion. It might be urged that the quus function is ruled out as an interpretation of '+' because it fails to satisfy some of the laws I accept for '+' (for example, it is not associative; we could have defined it so as not even to be commutative). One might even observe that, on the natural numbers, addition is the only function that satisfies certain laws that I accept – the 'recursion equations' for +: (x) $(x+0=x)$ and (x) (y) $(x+y'=(x+y)')$ where the stroke or dash indicates successor; these equations are sometimes called a 'definition' of addition. The problem is that the other signs used in these laws (the universal quantifiers, the equality sign) have been applied in only a finite number of instances, and they can be given non-standard interpretations that will fit non-standard interpretations of '+'. Thus for example '(x)' might mean for every $x<h$, where h is some upper bound to the instances where universal instantiation has hitherto been applied, and similarly for equality.

In any event the objection is somewhat overly sophisticated. Many of us who are not mathematicians use the '+' sign perfectly well in ignorance of any explicitly formulated laws of the type cited.

It is pointless of course to protest that I intended the result of counting a heap to be *independent* of its composition in terms of sub-heaps. Let me have said this to myself as explicitly as possible: the sceptic will smilingly reply that once again I am misinterpreting my past usage, that actually 'independent' formerly meant *quindependent*, where 'quindependent' means . . .

Here of course I am expounding Wittgenstein's well-known remarks about "a rule for interpreting a rule". It is tempting to answer the sceptic by appealing from one rule to another more 'basic' rule. But the sceptical move can be repeated at the more 'basic' level also. Eventually the process must stop – "justifications come to an end somewhere" – and I am left with a rule which is completely unreduced to any other. How can I justify my present application of such a rule, when a sceptic could easily interpret it so as to yield any of an indefinite number of other results? It seems that my application of it is an unjustified stab in the dark. I apply the rule *blindly*.

Normally, when we consider a mathematical rule such as addition, we think of ourselves as *guided* in our application of it to each new instance. Just this is the difference between someone who computes new values of a function and someone who calls out numbers at random. Given my past intentions regarding the symbol '+', one and only one answer

is dictated as the one appropriate to '68+57'. On the other hand, although an intelligence tester may suppose that there is only one possible continuation to the sequence 2, 4, 6, 8, . . ., mathematical and philosophical sophisticates know that an indefinite number of rules (even rules stated in terms of mathematical functions as conventional as ordinary polynomials) are compatible with any such finite initial segment. So if the tester urges me to respond, after 2, 4, 6, 8, . . ., with *the* unique appropriate next number, the proper response is that no such unique number exists, nor is there any unique (rule determined) infinite sequence that continues the given one. The problem can then be put this way: Did I myself, in the directions for the future that I gave myself regarding '+', really differ from the intelligence tester? True, I may not merely stipulate that '+' is to be a function instantiated by a finite number of computations. In addition, I may give myself directions for the further computation of '+', stated in terms of other functions and rules. In turn, I may give myself directions for the further computation of these functions and rules, and so on. Eventually, however, the process must stop, with 'ultimate' functions and rules that I have stipulated for myself only by a *finite* number of examples, just as in the intelligence test. If so, is not my procedure as arbitrary as that of the man who guesses the continuation of the intelligence test? In what sense is my actual computation procedure, following an algorithm that yields '125', more justified by my past instructions than an alternative procedure that would have resulted in '5'? Am I not simply following an unjustifiable impulse?[13]

[13] Few readers, I suppose, will by this time be tempted to appeal a determination to "go on the same way" as before. Indeed, I mention it at this point primarily to remove a possible misunderstanding of the sceptical argument, not to counter a possible reply to it. Some followers of Wittgenstein – perhaps occasionally Wittgenstein himself – have thought that his point involves a rejection of 'absolute identity' (as opposed to some kind of 'relative' identity). I do not see that this is so, whether or not doctrines of 'relative' identity are correct on other grounds. Let identity be as 'absolute' as one pleases: it holds only between each thing and itself. Then the plus function is identical with itself, and the quus function is identical with itself. None of this will tell me whether I referred to the plus function or to the quus function in the past, nor therefore will it tell me which to use in order to apply the same function now.

Wittgenstein does insist that the law of identity ('everything is identical with itself') gives no way out of this problem. It should be clear enough that this is so (whether or not the slogan should be rejected as 'useless'). Wittgenstein sometimes writes as if the way we give a response in a new case determines what we call the 'same', as if the meaning of 'same' varies from case to case. Whatever impression this gives, it need not relate to doctrines of relative and absolute identity. The point (which can be fully understood only after the third section of the present work) can be put this way: If someone who computed '+' as we do for small arguments gave bizarre responses, in the style of 'quus', for larger arguments, and insisted that he was 'going on the same way as before', we would not acknowledge his claim that he was 'going on in the same way' as for the small arguments. What we call the 'right' response determines what we call 'going on in the same way'. None of this in itself implies that identity is 'relative' in senses that 'relative identity' has been used elsewhere in the literature.

In fairness to Peter Geach, the leading advocate of the 'relativity' of identity, I should mention (lest the reader assume I had him in mind) that he is *not* one of those I have heard expound Wittgenstein's doctrine as dependent on a denial of 'absolute' identity.

Of course, these problems apply throughout language and are not confined to mathematical examples, though it is with mathematical examples that they can be most smoothly brought out. I think that I have learned the term 'table' in such a way that it will apply to indefinitely many future items. So I can apply the term to a new situation, say when I enter the Eiffel Tower for the first time and see a table at the base. Can I answer a sceptic who supposes that by 'table' in the past I meant *tabair*, where a 'tabair' is anything that is a table not found at the base of the Eiffel Tower, or a chair found there? Did I think explicitly of the Eiffel Tower when I first 'grasped the concept of' a table, gave myself directions for what I meant by 'table'? And even if I did think of the Tower, cannot any directions I gave myself mentioning it be reinterpreted compatibly with the sceptic's hypothesis? Most importantly

for the 'private language' argument, the point of course applies to predicates of sensations, visual impressions, and the like, as well: "*How do I know* that in working out the series + 2 I must write "20,004, 20,006" and not "20,004, 20,008"? – (The question: "How do I know that this color is 'red'?" is similar.)" (*Remarks on the Foundations of Mathematics*, I, §3) The passage strikingly illustrates a central thesis of this essay: that Wittgenstein regards the fundamental problems of the philosophy of mathematics and of the 'private language argument' – the problem of sensation language – as at root identical, stemming from his paradox. The whole of §3 is a succinct and beautiful statement of the Wittgensteinian paradox; indeed the whole initial section of part I of *Remarks on the Foundations of Mathematics* is a development of the problem with special reference to mathematics and logical inference. It has been supposed that all I need to do to determine my use of the word 'green' is to have an image, a sample, of green that I bring to mind whenever I apply the word in the future. When I use this to justify my application of 'green' to a new object, should not the sceptical problem be obvious to any reader of Goodman?[14] Perhaps by 'green', in the past I meant *grue*,[15] and the color image, which indeed was grue, was meant to direct me to apply the word 'green' to *grue* objects always. If the *blue* object before me now is grue, then it falls in the extension of 'green', as I meant it in the past. It is no help to suppose that in the past I stipulated that 'green' was to apply to all and only those things 'of the same color as' the sample. The sceptic can reinterpret 'same color' as same *schmolor*,[16] where things have the same schmolor if . . .

[14] See Nelson Goodman, *Fact, Fiction, and Forecast* (3rd ed., Bobbs-Merrill, Indianapolis, 1973, xiv+131 pp.).

[15] The exact definition of 'grue' is unimportant. It is best to suppose that past objects were grue if and only if they were (then) green while present objects are grue if and only if they are (now) blue. Strictly speaking, this is not Goodman's original idea, but it is probably most convenient for present purposes. Sometimes Goodman writes this way as well.

[16] 'Schmolor', with a slightly different spelling, appears in Joseph Ullian, "More on 'Grue' and Grue," *The Philosophical Review*, vol. 70 (1961), pp. 386–9.

Let us return to the example of 'plus' and 'quus'. We have just summarized the problem in terms of the basis of my present particular response: what tells me that I should say '125' and not '5'? Of course the problem can be put equivalently in terms of the sceptical query regarding my present intent: nothing in my mental history establishes whether I meant plus or quus. So formulated, the problem may appear to be epistemological – how can anyone know which of these I meant? Given, however, that everything in my mental history is compatible both with the conclusion that I meant plus and with the conclusion that I meant quus, it is clear that the sceptical challenge is not really an epistemological one. It purports to show that nothing in my mental history of past behavior – not even what an omniscient God would know – could establish whether I meant plus or quus. But then it appears to follow that there was no *fact* about me that constituted my having meant plus rather than quus. How could there be, if nothing in my internal mental history or external behavior will answer the sceptic who supposes that in fact I meant quus? If there was no such thing as my meaning plus rather than quus in the past, neither can there be any such thing in the present. When we initially presented the paradox, we perforce used language, taking present meanings for granted. Now we see, as we expected, that this provisional concession was indeed fictive. There can be no fact as to what I mean by 'plus', or any other word at any time. The ladder must finally be kicked away.

This, then, is the sceptical paradox. When I respond in one way rather than another to such a problem as '68+57', I can have no justification for one response rather than another. Since the sceptic who supposes that I meant quus cannot be answered, there is no fact about me that distinguishes between my meaning plus and my meaning quus. Indeed, there is no fact about me that distinguishes between my meaning a definite function by 'plus' (which determines my responses in new cases) and my meaning nothing at all.

Sometimes when I have contemplated the situation, I have had something of an eerie feeling. Even now as I write, I feel

confident that there is something in my mind – the meaning I attach to the 'plus' sign – that *instructs* me what I ought to do in all future cases. I do not *predict* what I *will* do – see the discussion immediately below – but instruct myself what I ought to do to conform to the meaning. (Were I now to make a prediction of my future behavior, it would have substantive content only because it already makes sense, in terms of the instructions I give myself, to ask whether my intentions will be conformed to or not.) But when I concentrate on what is now in my mind, what instructions can be found there? How can I be said to be acting on the basis of these instructions when I act in the future? The infinitely many cases of the table are not in my mind for my future self to consult. To say that there is a general rule in my mind that tells me how to add in the future is only to throw the problem back on to other rules that also seem to be given only in terms of finitely many cases. What can there be in my mind that I make use of when I act in the future? It seems that the entire idea of meaning vanishes into thin air.

Can we escape these incredible conclusions? Let me first discuss a response that I have heard more than once in conversation on this topic. According to this response, the fallacy in the argument that no fact about me constitutes my meaning plus lies in the assumption that such a fact must consist in an *occurent* mental state. Indeed the sceptical argument shows that my entire occurent past mental history might have been the same whether I meant plus or quus, but all this shows is that the fact that I meant plus (rather than quus) is to be analyzed *dispositionally*, rather than in terms of occurent mental states. Since Ryle's *The Concept of Mind*, dispositional analyses have been influential; Wittgenstein's own later work is of course one of the inspirations for such analyses, and some may think that he himself wishes to suggest a dispositional solution to his paradox.

The dispositional analysis I have heard proposed is simple. To mean addition by '+' is to be disposed, when asked for any sum '$x+y$' to give the sum of x and y as the answer (in

particular, to say '125' when queried about '68+57'); to mean quus is to be disposed when queried about any arguments, to respond with their *quum* (in particular to answer '5' when queried about '68+57'). True, my actual thoughts and responses in the past do not differentiate between the plus and the quus hypotheses; but, even in the past, there were dispositional facts about me that did make such a differentiation. To say that in fact I meant plus in the past is to say – as surely was the case! – that had I been queried about '68 + 57', I *would* have answered '125'. By hypothesis I was not in fact asked, but the disposition was present none the less.

To a good extent this reply immediately ought to appear to be misdirected, off target. For the sceptic created an air of puzzlement as to my *justification* for responding '125' rather than '5' to the addition problem as queried. He thinks my response is no better than a stab in the dark. Does the suggested reply advance matters? How does it *justify* my choice of '125'? What it says is: "'125' is the response you are disposed to give, and (perhaps the reply adds) it would also have been your response in the past." Well and good, I know that '125' is the response I am disposed to give (I am actually giving it!), and maybe it is helpful to be told – as a matter of brute fact – that I would have given the same response in the past. How does any of this indicate that – now *or* in the past – '125' was an answer *justified* in terms of instructions I gave myself, rather than a mere jack-in-the-box unjustified and arbitrary response? Am I supposed to justify my present belief that I meant addition, not quaddition, and hence should answer '125', in terms of a *hypothesis* about my *past* dispositions? (Do I record and investigate the past physiology of my brain?) Why am I so sure that one particular hypothesis of this kind is correct, when all my past thoughts can be construed either so that I meant plus or so that I meant quus? Alternatively, is the hypothesis to refer to my *present* dispositions alone, which would hence give the right answer by definition?

Nothing is more contrary to our ordinary view – or

Wittgenstein's – than is the supposition that "whatever is going to seem right to me is right." (§258). On the contrary, "that only means that here we can't talk about right" (*ibid.*). A candidate for what constitutes the state of my meaning one function, rather than another, by a given function sign, ought to be such that, whatever in fact I (am disposed to) do, there is a unique thing that I *should* do. Is not the dispositional view simply an equation of performance and correctness? Assuming determinism, even if I mean to denote *no* number theoretic function in particular by the sign '*', to the same extent as it is true for '+', it is true here that for any two arguments *m* and *n*, there is a uniquely determined answer *p* that I would give[17] (I choose one at random, as we would normally say, but causally the answer is determined). The difference between this case and the case of the '+' function is that in the former case, but not in the latter, my uniquely determined answer can properly be called 'right' or 'wrong'.[18]

So it does seem that a dispositional account misconceives the sceptic's problem – to find a past fact that *justifies* my present response. As a candidate for a 'fact' that determines what I mean, it fails to satisfy the basic condition on such a candidate, stressed above on p. 11, that it should *tell* me what I ought to do in each new instance. Ultimately, almost all objections to the dispositional accounts boil down to this one. However, since the dispositionalist does offer a popular

[17] We will see immediately below that for arbitrarily large *m* and *n*, this assertion is not really true even for '+'. That is why I say that the assertion is true for '+' and the meaningless '*' 'to the same extent'.

[18] I might have introduced '*' to mean nothing in particular even though the answer I arbitrarily choose for '*m*n*' is, through some quirk in my brain structure, uniquely determined independently of the time and other circumstances when I am asked the question. It might, in addition, even be the case that I consciously resolve, once I have chosen a particular answer to '*m*n*', to stick to it if the query is repeated for any particular case, yet nevertheless I think of '*' as meaning no function in particular. What I will not say is that my particular answer is 'right' or 'wrong' in terms of the *meaning* I assigned to '*', as I will for '+', since there is no such meaning.

candidate for what the fact as to what I mean might be, it is worth examining some problems with the view in more detail.

As I said, probably some have read Wittgenstein himself as favoring a dispositional analysis. I think that on the contrary, although Wittgenstein's views have dispositional elements, any such analysis is inconsistent with Wittgenstein's view.[19]

[19] Russell's *The Analysis of Mind* (George Allen and Unwin, London, in the Muirhead Library of Philosophy, 310 pp.) already gives dispositional analyses of certain mental concepts: see especially, Lecture III, "Desire and Feeling," pp. 58–76. (The object of a desire, for example, is roughly defined as that thing which, when obtained, will cause the activity of the subject due to the desire to cease.) The book is explicitly influenced by Watsonian behaviorism; see the preface and the first chapter. I am inclined to conjecture that Wittgenstein's philosophical development was influenced considerably by this work, both in the respects in which he sympathizes with behavioristic and dispositional views, and to the extent that he opposes them. I take *Philosophical Remarks* (Basil Blackwell, Oxford, 1975, 357 pp., translated by R. Hargreaves and R. White), §§21ff., to express a rejection of Russell's theory of desire, as stated in Lecture III of *The Analysis of Mind*. The discussion of Russell's theory played, I think, an important role in Wittgenstein's development: the problem of the relation of a desire, expectation, etc., to its object ('intentionality') is one of the important forms Wittgenstein's problem about meaning and rules takes in the *Investigations*. Clearly the sceptic, by proposing his bizarre interpretations of what I previously meant, can set bizarre results as to what (in the present) does, or does not, satisfy my past desires or expectations, or what constitutes obedience to an order I gave. Russell's theory parallels the dispositional theory of meaning in the text by giving a causal dispositional account of desire. Just as the dispositional theory holds that the value I meant '+' to have for two particular arguments *m* and *n* is, by definition, the answer I would give if queried about '*m*+*n*', so Russell characterizes the thing I desired as the thing which, were I to get it, would quiet my 'searching' activity. I think that even in the *Investigations*, as in *Philosophical Remarks* (which stems from an earlier period), Wittgenstein still rejects Russell's dispositional theory because it makes the relation between a desire and its object an 'external' relation (*PR*, §21), although in the *Investigations*, unlike *Philosophical Remarks*, he no longer bases this view on the 'picture theory' of the *Tractatus*. Wittgenstein's view that the relation between the desire (expectation, etc.) and its object must be 'internal', not 'external',

First, we must state the simple dispositional analysis. It gives a criterion that will tell me what number theoretic function φ I mean by a binary function symbol 'f', namely: The referent φ of 'f' is that unique binary function φ such that I am disposed, if queried about '$f(m, n)$', where 'm' and 'n' are numerals denoting particular numbers m and n, to reply 'p', where 'p' is a numeral denoting $\varphi(m, n)$. The criterion is meant to enable us to 'read off' which function I mean by a given function symbol from my disposition. The cases of addition and quaddition above would simply be special cases of such a scheme of definition.[20]

The dispositional theory attempts to avoid the problem of the finiteness of my actual past performance by appealing to a disposition. But in doing so, it ignores an obvious fact: not only my actual performance, but also the totality of my dispositions, is finite. It is not true, for example, that if queried about the sum of any two numbers, no matter how large, I will reply with their actual sum, for some pairs of numbers are

parallels corresponding morals drawn about meaning in my text below (the relation of meaning and intention to future action is 'normative, not descriptive', p. 37 below). Sections 429–65 discuss the fundamental problem of the *Investigations* in the form of 'intentionality'. I am inclined to take §440 and §460 to refer obliquely to Russell's theory and to reject it.

Wittgenstein's remarks on machines (see pp. 33–4 and note 24 below) also express an explicit rejection of dispositional and causal accounts of meaning and following a rule.

[20] Actually such a crude definition is quite obviously inapplicable to functions that I can define but cannot compute by any algorithm. Granted Church's thesis, such functions abound. (See the remark on Turing machines in footnote 24 below.) However, Wittgenstein himself does not consider such functions when he develops his paradox. For symbols denoting such functions the question "What function do I mean by the symbol?" makes sense; but the usual Wittgensteinian paradox (any response, not just the one I give, accords with the rule) makes no sense, since there need be no response that I give if I have no procedure for computing values of the function. Nor does a dispositional account of what I mean make sense. – This is not the place to go into such matters: for Wittgenstein, it may be connected with his relations to finitism and intuitionism.

simply too large for my mind – or my brain – to grasp. When given such sums, I may shrug my shoulders for lack of comprehension; I may even, if the numbers involved are large enough, die of old age before the questioner completes his question. Let 'quaddition' be redefined so as to be a function which agrees with addition for all pairs of numbers small enough for me to have any disposition to add them, and let it diverge from addition thereafter (say, it is 5). Then, just as the sceptic previously proposed the hypothesis that I meant quaddition in the old sense, now he proposes the hypothesis that I meant quaddition in the new sense. A dispositional account will be impotent to refute him. As before, there are infinitely many candidates the sceptic can propose for the role of quaddition.

I have heard it suggested that the trouble arises solely from too crude a notion of disposition: *ceteris paribus*, I surely will respond with the sum of any two numbers when queried. And *ceteris paribus* notions of dispositions, not crude and literal notions, are the ones standardly used in philosophy and in science. Perhaps, but how should we flesh out the *ceteris paribus* clause? Perhaps as something like: if my brain had been stuffed with sufficient extra matter to grasp large enough numbers, and if it were given enough capacity to perform such a large addition, and if my life (in a healthy state) were prolonged enough, then given an addition problem involving two large numbers, m and n, I would respond with their sum, and not with the result according to some quus-like rule. But how can we have any confidence of this? How in the world can I tell what would happen if my brain were stuffed with extra brain matter, or if my life were prolonged by some magic elixir? Surely such speculation should be left to science fiction writers and futurologists. We have no idea what the results of such experiments would be. They might lead me to go insane, even to behave according to a quus-like rule. The outcome really is obviously indeterminate, failing further specification of these magic mind-expanding processes; and even with such specifications, it is highly speculative. But of course what the

ceteris paribus clause really means is something like this: If I somehow were to be given the means to carry out my intentions with respect to numbers that presently are too long for me to add (or to grasp), and if I were to carry out these intentions, then if queried about '$m+n$' for some big m and n, I would respond with their sum (and not with their quum). Such a counterfactual conditional is true enough, but it is of no help against the sceptic. It presupposes a prior notion of my having an intention to mean one function rather than another by '+'. It is in virtue of a fact of this kind about me that the conditional is true. But of course the sceptic is challenging the existence of just such a fact; his challenge must be met by specifying its nature. Granted that I mean addition by '+', then of course if I were to act in accordance with my intentions, I would respond, given any pair of numbers to be combined by '+', with their sum; but equally, granted that I mean quaddition, if I were to act in accordance with my intentions, I would respond with the quum. One cannot favor one conditional rather than another without circularity.

Recapitulating briefly: if the dispositionalist attempts to define which function I meant as the function determined by the answer I am disposed to give for arbitrarily large arguments, he ignores the fact that my dispositions extend to only finitely many cases. If he tries to appeal to my responses under idealized conditions that overcome this finiteness, he will succeed only if the idealization includes a specification that I will still respond, under these idealized conditions, according to the infinite table of the function I actually meant. But then the circularity of the procedure is evident. The idealized dispositions are determinate only because it is already settled which function I meant.

The dispositionalist labors under yet another, equally potent, difficulty, which was foreshadowed above when I recalled Wittgenstein's remark that, if 'right' makes sense, it cannot be the case that whatever seems right to me is (by definition) right. Most of us have dispositions to make

mistakes.[21] For example, when asked to add certain numbers some people forget to 'carry'. They are thus disposed, for these numbers, to give an answer differing from the usual addition table. Normally, we say that such people have made a *mistake*. That means, that for them as for us, '+' means addition, but for certain numbers they are not disposed to give the answer they *should* give, if they are to accord with the table of the function they actually *meant*. But the dispositionalist cannot say this. According to him, the function someone means is to be *read off* from his dispositions; it cannot be

[21] However, in the slogan quoted and in §202, Wittgenstein seems to be more concerned with the question, "Am I right in thinking that I am still applying the same rule?" than with the question "Is my application of the rule right?" Relatively few of us have the disposition – as far as I know – bizarrely to cease to apply a given rule if once we were applying it. Perhaps there is a corrosive substance present in my brain already (whose action will be 'triggered' if I am given a certain addition problems) that will lead me to forget how to add. I might, once this substance is secreted, start giving bizarre answers to addition problems – answers that conform to a quus-like rule, or to no discernible pattern at all. Even if I do think that I am following the same rule, in fact I am not.

Now, when I assert that I definitely mean addition by 'plus', am I making a *prediction* about my future behavior, asserting that there is no such corrosive acid? To put the matter differently: I assert that the present meaning I give to '+' determines values for arbitrarily large amounts. I do *not* predict that I will come out with these values, or even that I will use anything like the 'right' procedures to get them. A disposition to go beserk, to change the rule, etc., may be in me already, waiting to be triggered by the right stimulus. I make no assertion about such possibilities when I say that my use of the '+' sign determines values for every pair of arguments. Much less do I assert that the values I will come out with under these circumstances are, by definition, the values that accord with what is meant.

These possibilities, and the case mentioned above with '*', when I am disposed to respond even though I follow no rule from the beginning, should be borne in mind in addition to the garden-variety possibility of error mentioned in the text. Note that in the case of '*', it seems intuitively possible that I could be under the impression that I was following a rule even though I was following none – see the analogous case of reading on pp. 45–6 below, in reference to §166.

presupposed in advance which function is meant. In the present instance a certain unique function (call it 'skaddition') corresponds in its table exactly to the subject's dispositions including his dispositions to make mistakes. (Waive the difficulty that the subject's dispositions are finite: suppose he has a disposition to respond to any pair of arguments.) So, where common sense holds that the subject means the same addition function as everyone else but systematically makes computational mistakes, the dispositionalist seems forced to hold that the subject makes no computational mistakes, but means a non-standard function ('skaddition') by '+'. Recall that the dispositionalist held that we would detect someone who meant quus by '+' *via* his disposition to respond with '5' for arguments $\geqslant 57$. I the same way, he will 'detect' that a quite ordinary, though fallible, subject means some non-standard function by '+'.

Once again, the difficulty cannot be surmounted by a *ceteris paribus* clause, by a clause excluding 'noise', or by a distinction between 'competence' and 'performance'. No doubt a disposition to give the true sum in response to each addition problem is part of my 'competence', if by this we mean simply that such an answer accords with the rule I intended, or if we mean that, if all my dispositions to make mistakes were removed, I would give the correct answer. (Again I waive the finiteness of my capacity.) But a disposition to make a mistake is simply a disposition to *give an answer other than the one that accords with the function I meant*. To presuppose this concept in the present discussion is of course viciously circular. If I meant addition, my 'erroneous' actual disposition is to be ignored; if I meant skaddition, it should not be. Nothing in the notion of my 'competence' as thus defined can possibly tell me which alternative to adopt.[22] Alternatively, we might try to specify

[22] Lest I be misunderstood, I hope it is clear that in saying this I do not myself reject Chomsky's competence–performance distinction. On the contrary, I personally find that the familiar arguments for the distinction (and for the attendant notion of grammatical rule) have great persuasive force. The present work is intended to expound my understanding of

the 'noise' to be ignored without presupposing a prior notion of which function is meant. A little experimentation will reveal the futility of such an effort. Recall that the subject has a

---

Wittgenstein's position, not my own; but I certainly do not mean, exegetically, to assert that Wittgenstein himself would reject the distinction. But what *is* important here is that the notion of 'competence' is itself not a dispositional notion. It is normative, not descriptive, in the sense explained in the text.

The point is that our understanding of the notion of 'competence' is dependent on our understanding of the idea of 'following a rule', as is argued in the discussion above. Wittgenstein would reject the idea that 'competence' can be defined in terms of an idealized dispositional or mechanical model, and used without circularity to explicate the notion of following a rule. Only after the sceptical problem about rules has been resolved can we *then* define 'competence' in terms of rule-following. Although notions of 'competence' and 'performance' differ (at least) from writer to writer, I see no reason why linguists need assume that 'competence' is defined prior to rule-following. Although the remarks in the text warn against the use of the 'competence' notion as a solution to our problem, in no way are they arguments against the notion itself.

Nevertheless, given the sceptical nature of Wittgenstein's solution to his problem (as this solution is explained below), it is clear that if Wittgenstein's standpoint is accepted, the notion of 'competence' will be seen in a light radically different from the way it implicitly is seen in much of the literature of linguistics. For *if* statements attributing rule-following are neither to be regarded as stating facts, nor to be thought of as *explaining* our behavior (see section 3 below), it would seem that the *use* of the ideas of rules and of competence in linguistics needs serious reconsideration, even if these notions are not rendered 'meaningless'. (Depending on one's standpoint, one might view the tension revealed here between modern linguistics and Wittgenstein's sceptical critique as casting doubt on the linguistics, or on Wittgenstein's sceptical critique – or both.) These questions would arise even if, as throughout the present text, we deal with rules, like addition, that are stated explicitly. These rules we think of ourselves as grasping consciously; in the absence of Wittgenstein's sceptical arguments, we would see no problem in the assumption that each particular answer we produce is justified by our 'grasp' of the rules. The problems are compounded if, as in linguistics, the rules are thought of as tacit, to be reconstructed by the scientist and *inferred* as an *explanation* of behavior. The matter deserves an extended discussion elsewhere. (See also p. 37 below.)

systematic disposition to forget to carry in certain circumstances: he tends to give a uniformly erroneous answer when well rested, in a pleasant environment free of clutter, etc. One cannot repair matters by urging that the subject would eventually respond with the right answer after correction by others. First, there are uneducable subjects who will persist in their error even after persistent correction. Second, what is meant by 'correction by others'? If it means rejection by others of 'wrong' answers (answers that do not accord with the rule the speaker means) and suggestion of the right answer (the answer that does accord), then again the account is circular. If random intervention is allowed (that is, the 'corrections' may be arbitrary, whether they are 'right' or 'wrong'), then, although educable subjects may be induced to correct their wrong answers, suggestible subjects may also be induced to replace their correct answers with erroneous ones. The amended dispositional statement will, then, provide no criterion for the function that is really meant.

The dispositional theory, as stated, assumes that which function I meant is determined by my dispositions to compute its values in particular cases. In fact, this is not so. Since dispositions cover only a finite segment of the total function and since they may deviate from its true values, two individuals may agree on their computations in particular cases even though they are actually computing different functions. Hence the dispositional view is not correct.

In discussions, I have sometimes heard a variant of the dispositional account. The argument goes as follows: the sceptic argues, in essence, that I am free to give any new answer to an addition problem, since I can always interpret my previous intentions appropriately. But how can this be? As Dummett put the objection: "A machine can follow this rule; whence does a human being gain a freedom of choice in this matter which a machine does not possess?"[23] The objection is

[23] M. A. E. Dummett, "Wittgenstein's Philosophy of Mathematics," *The Philosophical Review*, vol. 68 (1959), pp. 324–48, see p. 331, reprinted in George Pitcher (ed.), *Wittgenstein: The Philosophical Investigations* (Macmillan, 1966, pp. 420–47), see p. 428. The quoted objection need not necessarily be taken to express Dummett's own ultimate view of the matter.

really a form of the dispositional account, for that account can be viewed as if it interpreted us as machines, whose output mechanically yields the correct result.

We can interpret the objector as arguing that the rule can be *embodied* in a machine that computes the relevant function. If I build such a machine, it will simply grind out the right answer, in any particular case, to any particular addition problem. The answer that the machine would give is, then, the answer that I intended.

The term 'machine' is here, as often elsewhere in philosophy, ambiguous. Few of us are in a position to build a machine or draw up a program to embody our intentions; and if a technician performs the task for me, the sceptic can ask legitimately whether the technician has performed his task correctly. Suppose, however, that I am fortunate enough to be such an expert that I have the technical facility required to embody my own intentions in a computing machine, and I state that the machine is *definitive* of my own intentions. Now the word 'machine' here may refer to any one of various things. It may refer to a machine *program* that I draw up, embodying my intentions as to the operation of the machine. Then exactly the same problems arise for the program as for the original symbol '+': the sceptic can feign to believe that the program, too, ought to be interpreted in a quus-like manner. To say that a program is not something that I wrote down on paper, but an abstract mathematical object, gets us no further. The problem then simply takes the form of the question: what program (in the sense of abstract mathematical object) corresponds to the 'program' I have written on paper (in accordance with the way I meant it)? ('Machine' often seems to mean a program in one of these senses: a Turing 'machine', for example, would be better called a 'Turing program'.) Finally, however, I may build a concrete machine, made of metal and

gears (or transistors and wires), and declare that it embodies the function I intend by '+': the values that it gives are the values of the function I intend. However, there are several problems with this. First, even if I say that the machine embodies the function in this sense, I must do so in terms of instructions (machine 'language', coding devices) that tell me how to interpret the machine; further, I must declare explicitly that the function always takes values as given, in accordance with the chosen code, by the machine. But then the sceptic is free to interpret all these instructions in a non-standard, 'quus-like' way. Waiving this problem, there are two others – here is where the previous discussion of the dispositional view comes in. I cannot really insist that the values of the function are given by the machine. First, the machine is a finite object, accepting only finitely many numbers as input and yielding only finitely many as output – others are simply too big. Indefinitely many programs extend the actual finite behavior of the machine. Usually this is ignored because the designer of the machine intended it to fulfill just one program, but in the present context such an approach to the intentions of the designer simply gives the sceptic his wedge to interpret in a non-standard way. (Indeed, the appeal to the designer's program makes the physical machine superfluous; only the program is really relevant. The machine as physical object is of value only if the intended function can somehow be read off from the physical object alone.) Second, in practice it hardly is likely that I really intend to entrust the values of a function to the operation of a physical machine, even for that finite portion of the function for which the machine can operate. Actual machines can *malfunction*: through melting wires or slipping gears they may give the wrong answer. How is it determined when a malfunction occurs? By reference to the program of the machine, as intended by its designer, not simply by reference to the machine itself. Depending on the intent of the designer, any particular phenomenon may or may not count as a machine 'malfunction'. A programmer with suitable intentions might even have intended to make use

of the fact that wires melt or gears slip, so that a machine that is 'malfunctioning' for me is behaving perfectly for him. Whether a machine ever malfunctions and, if so, when, is not a property of the machine itself as a physical object but is well defined only in terms of its program, as stipulated by its designer. Given the program, once again the physical object is superfluous for the purpose of determining what function is meant. Then, as before, the sceptic can concentrate his objections on the program. The last two criticisms of the use of the physical machine as a way out of scepticism – its finitude and the possibility of malfunction – obviously parallel two corresponding objections to the dispositional account.[24]

[24] Wittgenstein discusses machines explicitly in §§193–5. See the parallel discussion in *Remarks on the Foundations of Mathematics*, part I, §§118–30, especially §§119–25; see also, e.g., II [III], §87, and III [IV], §§48–9 there. The criticisms in the text of the dispositional analysis and of the use of machines to solve the problem are inspired by these sections. In particular, Wittgenstein himself draws the distinction between the machine as an abstract program ("der Maschine, als Symbol" §193) and the actual physical machine, which is subject to breakdown ("do we forget the possibility of their bending, breaking off, melting, and so on?" (§193)). The dispositional theory views the subject himself as a kind of machine, whose potential actions embody the function. So in this sense the dispositional theory and the idea of the machine-as-embodying-the-function are really one. Wittgenstein's attitude toward both is the same: they confuse the 'hardness of a rule' with the 'hardness of a material' (*RFM*, II [III], §87). On my interpretation, then, Wittgenstein agrees with his interlocutor (§194 and §195) that the sense in which all the values of the function are already present is not simply causal, although he disagrees with the idea that the future use is already present in some mysterious non-causal way.

Although, in an attempt to follow Wittgenstein, I have emphasized the distinction between concrete physical machines and their abstract programs in what I have written above, it might be instructive to look at the outcome when the limitation of machines is idealized as in the modern theory of automata. A finite automaton, as usually defined, has only finitely many states, receives only finitely many distinct inputs, and has only finitely many outputs, but it is idealized in two respects: it has no problem of malfunction, and its lifetime (without any decay or wearing out of its parts) is infinite. Such a machine can, in a sense, perform computations on arbitrarily large whole numbers. If it has notations for

the single digits from zero through nine, inclusive, it can receive arbitrarily large positive whole numbers as inputs simply by being given their digits one by one. (We cannot do this, since our effective lifetimes are finite, and there is a minimum time needed for us to understand any single digit.) Such an automaton can add according to the usual algorithm in decimal notation (the digits for the numbers being added should be fed into the machine starting from the last digits of both summands and going backwards, as in the usual algorithm). However, it can be proved that, in the same ordinary decimal notation, such a machine cannot *multiply*. Any function computed by such a machine that purports to be multiplication will, for large enough arguments, exhibit 'quus-like' (or rather, 'quimes-like') properties at sufficiently large arguments. Even if we were idealized as finite automata, a dispositional theory would yield unacceptable results.

Suppose we idealized even further and considered a Turing machine which has a tape to use which is infinite in both directions. Such a machine has infinite extent at every moment, in addition to an infinite lifetime without malfunctions. Turing machines can multiply correctly, but it is well known that even here there are many functions we can define explicitly that can be computed by no such machine. A crude dispositional theory would attribute to us a non-standard interpretation (or no interpretation at all(for any such function. (See above, note 20.)

I have found that both the crude dispositional theory and the function-as-embodied-in-a-machine come up frequently when Wittgenstein's paradox is discussed. For this reason, and because of their close relation to Wittgenstein's text, I have expounded these theories, though sometimes I have wondered whether the discussion of them is excessively long. On the other hand, I have resisted the temptation to discuss 'functionalism' explicitly, even though various forms of it have been so attractive to so many of the best recent writers that it has almost become the received philosophy of mind in the USA. Especially I have feared that some readers of the discussion in the text will think that 'functionalism' is precisely the way to modify the crude dispositional theory so as to meet the criticisms (especially those that rely on the circularity of *ceteris paribus* clauses). (I report, however, that thus far I have not run into such reactions in practice.) I cannot discuss functionalism at length here without straying from the main point. But I offer a brief hint. Functionalists are fond of comparing psychological states to the abstract states of a (Turing) machine, though some are cognizant of certain limitations of the comparison. All regard psychology as given by a set of causal connections, analogous to the *causal* operation of a machine. But then the remarks of the text stand here as well: any concrete physical object can be viewed as an imperfect realization of many machine programs. Taking a human organism as a concrete object, what is to tell

The moral of the present discussion of the dispositional account may be relevant to other areas of concern to philosophers beyond the immediate point at issue. Suppose I do mean addition by '+'. What is the relation of this supposition to the question how I will respond to the problem '68+57'? The dispositionalist gives a *descriptive* account of this relation: if '+' meant addition, then I will answer '125'. But this is not the proper account of the relation, which is *normative*, not descriptive. The point is *not* that, if I meant addition by '+', I *will* answer '125', but that, if I intend to accord with my past meaning of '+', I *should* answer '125'. Computational error, finiteness of my capacity, and other disturbing factors may lead me not to be *disposed* to respond as I *should*, but if so, I have not acted in accordance with my intentions. The relation of meaning and intention to future action is *normative*, not *descriptive*.

In the beginning of our discussion of the dispositional analysis, we suggested that it had a certain air of irrelevance with respect to a significant aspect of the sceptical problem – that the fact that the sceptic can maintain the hypothesis that I meant quus shows that I had no *justification* for answering '125' rather than '5'. How does the dispositional analysis even appear to touch this problem? Our conclusion in the previous paragraph shows that in some sense, after giving a number of more specific criticisms of the dispositional theory, we have returned full circle to our original intuition. Precisely the fact that our answer to the question of which function I meant is *justificatory* of my present response is ignored in the dispositional account and leads to all its difficulties.

I shall leave the dispositional view. Perhaps I have already belabored it too much. Let us repudiate briefly another

---

us *which* program he should be regarded as instantiating? In particular, does he compute 'plus' or 'quus'? If the remarks on machines in my own (and Wittgenstein's) text are understood, I think it will emerge that as far as the present problem is concerned, Wittgenstein would regard his remarks on machines as applicable to 'functionalism' as well.

I hope to elaborate on these remarks elsewhere.

suggestion. Let no one – under the influence of too much philosophy of science – suggest that the hypothesis that I meant plus is to be preferred as the *simplest* hypothesis. I will not here argue that simplicity is relative, or that it is hard to define, or that a Martian might find the quus function simpler than the plus function. Such replies may have considerable merit, but the real trouble with the appeal to simplicity is more basic. Such an appeal must be based either on a misunderstanding of the sceptical problem, or of the role of simplicity considerations, or both. Recall that the sceptical problem was not merely epistemic. The sceptic argues that there is no fact as to what I meant, whether plus or quus. Now simplicity considerations can help us decide between competing hypotheses, but they obviously can never tell us what the competing hypotheses are. If we do not understand what two hypotheses *state*, what does it mean to say that one is 'more probable' because it is 'simpler'? If the two competing hypotheses are not genuine hypotheses, not assertions of genuine matters of fact, no 'simplicity' considerations will make them so.

Suppose there are two conflicting hypotheses about electrons, both confirmed by the experimental data. If our own view of statements about electrons is 'realist' and not 'instrumentalist', we will view these assertions as making factual assertions about some 'reality' about electrons. God, or some appropriate being who could 'see' the facts about electrons directly, would have no need for experimental evidence or simplicity considerations to decide between hypotheses. We, who lack such capacities, must rely on indirect evidence, from the effects of the electrons on the behavior of gross objects, to decide between the hypotheses. If two competing hypotheses are indistinguishable as far as their effects on gross objects are concerned, then *we* must fall back on simplicity considerations to decide between them. A being – not ourselves – who could 'see' the facts about electrons 'directly' would have no need to invoke simplicity considerations, nor to rely on indirect evidence to decide between the hypotheses; he would 'directly perceive' the relevant facts that

make one hypothesis true rather than another. To say this is simply to repeat, in colorful terminology, the assertion that the two hypotheses do state genuinely different matters of fact.

Now Wittgenstein's sceptic argues that he knows of no fact about an individual that could constitute his state of meaning plus rather than quus. Against *this* claim simplicity considerations are irrelevant. Simplicity considerations would have been relevant against a sceptic who argued that the indirectness of our access to the facts of meaning and intention *prevents us ever from knowing* whether we mean plus or quus. But such merely epistemological scepticism is *not* in question. The sceptic does not argue that our own limitations of access to the facts prevent us from knowing something hidden. He claims that an omniscient being, with access to *all* available facts, still would not find any fact that differentiates between the plus and the quus hypotheses. Such an omniscient being would have neither need nor use for simplicity considerations.[25]

[25] A different use of 'simplicity', not that by which we evaluate competing theories, might suggest itself with respect to the discussion of machines above. There I remarked that a concrete physical machine, considered as an object without reference to a designer, may (approximately) instantiate any number of programs that (approximately, allowing for some 'malfunctioning') extend its actual finite behavior. If the physical machine was not designed but, so to speak, 'fell from the sky', there can be no fact of the matter as to which program it 'really' instantiates, hence no 'simplest hypothesis' about this non-existent fact.

Nevertheless, given a physical machine, one might ask what is the *simplest program* that the physical machine approximates. To do this one would have to find a measure of the simplicity of programs, a measure of the trade-off of the simplicity of the program with the degree to which the concrete machine fails to conform to it (malfunctions), and so on. I who am no expert, nor even an amateur, am unaware that this problem has been considered by theoretical computer scientists. Whether or not it has been considered, intuition suggests that something might be made of it, though it would not be trivial to find simplicity measures that give intuitively satisfying results.

I doubt that any of this would illuminate Wittgenstein's sceptical paradox. One might try, say, to define the function I meant as the one that, according to the simplicity measure, followed the simplest program

The idea that we lack 'direct' access to the facts whether we mean plus or quus is bizarre in any case. Do I not know, directly, and with a fair degree of certainty, that I mean plus? Recall that a fact as to what I mean now is supposed to *justify* my future actions, to make them *inevitable* if I wish to use words with the same meaning with which I used them before. This was our fundamental requirement on a fact as to what I meant. No 'hypothetical' state could satisfy such a requirement: If I can only form hypotheses as to whether I now mean plus or quus, if the truth of the matter is buried deep in my unconscious and can only be posited as a tentative hypothesis, then in the future I can only proceed hestitatingly and hypothetically, *conjecturing* that I probably ought to answer '68+57' with '125' rather than '5'. Obviously, this is not an accurate account of the matter. There may be some facts about me to which my access is indirect, and about which I must form tentative hypotheses: but surely the fact as to what I mean by 'plus' is not one of them! To say that it is, is already to take a big step in the direction of scepticism. Remember that I immediately and unhesitatingly calculate '68 + 57' as I do, and the meaning I assign to '+' is supposed to *justify* this procedure. I do not form tentative hypotheses, wondering what I should do if one hypothesis or another were true.

Now the reference, in our exposition, to what an omniscient being could or would know is merely a dramatic device. When the sceptic denies that even God, who knows all the

approximately compatible with my physical structure. Suppose brain physiologists found – to their surprise – that actually such a simplicity measure led to a program that did not compute addition for the '+' function, but some other function. Would this show that I did not mean addition by '+'? Yet, in the absence of detailed knowledge of the brain (and the hypothetical simplicity measure), the physiological discovery in question is by no means inconceivable. The justificatory aspect of the sceptic's problem is even more obviously remote from any such simplicity measure. I do not justify my choice of '125' rather than '5' as an answer to '68+57' by citing a hypothetical simplicity measure of the type mentioned. (I hope to elaborate on this in the projected work on functionalism mentioned in note 24 above.)

facts, could know whether I meant plus or quus, he is simply giving colorful expression to his denial that there is any fact of the matter as to which I meant. Perhaps if we remove the metaphor we may do better. The metaphor, perhaps, may seduce us towards scepticism by encouraging us to look for a reduction of the notions of meaning and intention to something else. Why not argue that "meaning addition by 'plus'" denotes an irreducible experience, with its own special *quale*, known directly to each of us by introspection? (Headaches, tickles, nausea are examples of inner states with such *qualia*.)[26] Perhaps the "decisive move in the conjuring trick" has been made when the sceptic notes that I have performed only finitely many additions and challenges me, in the light of *this* fact, to adduce some fact that 'shows' that I did not mean quus. Maybe I appear to be unable to reply just because the experience of meaning addition by 'plus' is as unique and irreducible as that of seeing yellow or feeling a headache, while the sceptic's challenge invites me to look for another fact or experience to which this can be reduced.

I referred to an *introspectible* experience because, since each of us knows immediately and with fair certainty that he means addition by 'plus', presumably the view in question assumes we know this in the same way we know that we have headaches – by attending to the 'qualitative' character of our own experiences. Presumably the experience of *meaning addition* has its own irreducible quality, as does that of feeling a headache. The fact that I mean addition by 'plus' is to be identified with my possession of an experience of this quality.

Once again, as in the case of the dispositional account, the proffered theory seems to be off target as an answer to the original challenge of the sceptic. The sceptic wanted to know why I was so sure that I ought to say '125', when asked about '68+57'. I had never thought of this particular addition before: is not an interpretation of the '+' sign as quus compatible with everything I thought? Well, suppose I do in fact feel a certain

[26] It is well known that this type of view is characteristic of Hume's philosophy. See note 51 below.

headache with a very special quality whenever I think of the '+' sign. How on earth would this headache help me figure out whether I ought to answer '125' or '5' when asked about '68+57'? If I think the headache indicates that I ought to say '125', would there be anything about it to refute a sceptic's contention that, on the contrary, it indicates that I should say '5'? The idea that each of my inner states – including, presumably, meaning what I do by 'plus' – has its special discernible quality like a headache, a tickle, or the experience of a blue after-image, is indeed one of the cornerstones of classical empiricism. Cornerstone it may be, but it is very hard to see how the alleged introspectible *quale* could be relevant to the problem at hand.

Similar remarks apply even to those cases where the classical empiricist picture might seem to have a greater plausibility. This picture suggested that association of an image with a word (paradigmatically a visual one) determined its meaning. For example (§134), a drawing of a cube comes to my mind whenever I hear or say the word 'cube'. It should be obvious that this need not be the case. Many of us use words such as 'cube' even though no such drawing or image comes to mind. Let us suppose, however, for the moment that one does. 'In what sense can this picture fit or fail to fit a use of the word 'cube'"? – Perhaps you say: "It's quite simple; – if that picture comes to me and I point to a triangular prism for instance, and say it's a cube, then this use of the word doesn't fit the picture." But doesn't it fit? I have purposely so chosen the example that it is quite easy to imagine a *method of projection* according to which the picture does fit after all. The picture of the cube did indeed *suggest* a certain use to us, but it was possible for me to use it differently.' The sceptic could suggest that the image be used in non-standard ways. 'Suppose, however, that not merely the picture of the cube, but also the method of projection comes before our mind? – How am I to imagine this? – Perhaps I see before me a schema showing the method of projection: say a picture of two cubes connected by lines of projection. – But does this really get me any further?

Can't I now imagine different applications of this schema too?" (§141). Once again, a rule for interpreting a rule. No internal impression, with a *quale*, could possibly tell me in itself how it is to be applied in future cases. Nor can any pile up of such impressions, thought of as rules for interpreting rules, do the job.[27] The answer to the sceptic's problem, "What tells me how I am to apply a given rule in a new case?", must come from something outside any images or 'qualitative' mental states. This is obvious, in the case of 'plus' – it is clear enough that no internal state such as a headache, a tickle, an image, could do the job. (Obviously I do not have an image of the infinite table of the 'plus' function in my mind. Some such image would be the only candidate that even has surface plausibility as a device for telling me how to apply 'plus'.) It may be less obvious in other cases, such as 'cube', but in fact it is also true of such cases as well.

So: If there were a special experience of 'meaning' addition by 'plus', analogous to a headache, it would not have the properties that a state of meaning addition by 'plus' ought to have – it would not tell me what to do in new cases. In fact, however, Wittgenstein extensively argues in addition that the supposed unique special experience of meaning (addition by 'plus', etc.) does not exist. His investigation here is an introspective one, designed to show that the supposed unique experience is a chimera. Of all the replies to the sceptic he combats, the view of meaning as an introspectible experience is probably the most natural and fundamental. But for the present day audience I dealt with it neither first nor at greatest length, for, though the Humean picture of an irreducible 'impression' corresponding to each psychological state or event has tempted many in the past, it tempts relatively few today. In fact, if in the past it was too readily and simplistically assumed, at present its force is – at least in my personal opinion – probably too *little* felt. There are several reasons for this. One is that, in this instance, Wittgenstein's critique of alternative

[27] The remarks above, p. 20, on the use of an image, or even a physical sample, of green, make the same point.

views has been relatively well received and absorbed. And related writers – such as Ryle – have reinforced the critique of the Cartesian and Humean pictures. Another reason – unattractive to the present writer – has been the popularity of materialistic-behavioristic views that ignore the problem of felt qualities of mental states altogether, or at least attempt to analyze all such states away in broadly behavioristic terms.[28]

It is important to repeat in the present connection what I have said above: Wittgenstein does not base his considerations on any behavioristic *premise* that dismisses the 'inner'. On the contrary, much of his argumentation consists in detailed introspective considerations. Careful consideration of our inner lives, he argues, will show that there is no special inner experience of 'meaning' of the kind supposed by his opponent. The case is specifically in *contrast* with feeling a pain, seeing red, and the like.

It takes relatively little introspective acuteness to realize the dubiousness of the attribution of a special qualitative character to the 'experience' of meaning addition by 'plus'. Attend to what happened when I first learned to add. First, there may or may not have been a specifiable time, probably in my childhood, at which I suddenly felt (*Eureka!*) that I had grasped the rule for addition. If there was not, it is very hard to see in what the suppositious special experience of my learning to add consisted. Even if there was a particular time at which I could have shouted "*Eureka!*" – surely the exceptional case – in what did the attendant experience consist? Probably consideration of a few particular cases and a thought – "Now I've got it!" – or the like. Could just *this* be the content of an experience of 'meaning addition'? How would it have been different if I had

[28] Although there are clear classical senses of behaviorism in which such current philosophies of mind as 'functionalism' are not behaviorist, nevertheless, speaking for myself, I find much contemporary 'functionalism' (especially those versions that attempt to give 'functional' *analyses* of mental terms) is far too behavioristic for my own taste. It would require an extensive digression to go into the matter further here.

meant quus? Suppose I perform a particular addition now, say '5+7'. Is there any special quality to the experience? Would it have been different if I had been trained in, and performed, the corresponding quaddition? How different indeed would the *experience* have been if I had performed the corresponding multiplication ('5×7'), other than that I would have responded automatically with a different answer? (Try the experiment yourself.)

Wittgenstein returns to points like these repeatedly throughout *Philosophical Investigations*. In the sections where he discusses his sceptical paradox (§§137–242), after a general consideration of the alleged introspectible process of understanding, he considers the issue in connection with the special case of *reading* (§§156–78). By 'reading' Wittgenstein means reading out loud what is written or printed and similar activities: he is not concerned with understanding what is written. I myself, like many of my coreligionists, first learned to 'read' Hebrew in this sense before I could understand more than a few words of the language. Reading in this sense is a simple case of 'following a rule'. Wittgenstein points out that a beginner, who reads by laboriously spelling words out, may have an introspectible experience when he really reads, as opposed to pretending to 'read' a passage he has actually memorized in advance; but an experienced reader simply calls the words out and is aware of no special conscious experience of 'deriving' the words from the page. The experienced reader may 'feel' nothing different when he reads from what the beginner feels, or does not feel, when he pretends. And suppose a teacher is teaching a number of beginners to read. Some pretend, others occasionally get it right by accident, others have already learned to read. When has someone passed into the latter class? In general, there will not be an identifiable moment when this has happened: the teacher will judge of a given pupil that he has 'learned to read' if he passes tests for reading often enough. There may or may not be an identifiable moment when the pupil first *felt*, "Now I am reading!" but the

presence of such an experience is neither a necessary nor a sufficient condition for the teacher to judge of him that he is reading.

Again (§160), someone may, under the influence of a drug, or in a dream, be presented with a made-up 'alphabet' and utter certain words, with all the characteristic 'feeling' of reading, to the extent that such a 'feeling' exists at all. If, after the drug wears off (or he wakes up), he himself thinks he was uttering words at random with no real connection with the script, should we really say he was reading? Or, on the other hand, what if the drug leads him to read fluently from a genuine text, but with the 'sensation' of reciting something learned by heart? Wasn't he still reading?

It is by examples like these – *Philosophical Investigations* contains a wealth of examples and mental thought experiments beyond what I have summarized – that Wittgenstein argues that the supposed special 'experiences' associated with rule following are chimerical.[29] As I said, my own discussion

[29] The point should not be overstated. Although Wittgenstein does deny that there is any particular 'qualitative' experience like a headache, present when and only when we use a word with a certain meaning (or read, or understand, etc.), he does acknowledge a certain 'feel' to our meaningful use of a word that may under certain circumstances be lost. Many have had a fairly common experience: by repeating a word or phrase again and again, one may be able to deprive it of its normal 'life', so that it comes to sound strange and foreign, even though one is still able to utter it under the right circumstances. Here there is a special feeling of foreignness in a particular case. Could there be someone who always used words like a mechanism, without any 'feeling' of a distinction between this mechanistic type of use and the normal case? Wittgenstein is concerned with these matters in the second part of the *Investigations*, in connection with his discussion of 'seeing as' (section xi, pp. 193–229). Consider especially his remarks on 'aspect blindness', pp. 213–14, and the relation of 'seeing an aspect' to 'experiencing the meaning of a word', p. 214. (See his examples on p. 214: "What would you be missing . . . if you did not feel that a word lost its meaning and became a mere sound if it was repeated ten times over? . . . Suppose I had agreed on a code with someone; "tower" means bank. I tell him "Now go to the tower" – he understands me and acts accordingly, but he feels the word "tower" to be strange in this use, it has not yet 'taken on' the meaning." He gives many examples on pp. 213–18.)

Compare (as Wittgenstein does) the feeling of meaning a word as such-and-such (think of 'ill' now as verb, now as a noun, etc.), with the idea of visual aspects discussed at length in section xi of the second part of the *Investigations*. We can see the duck-rabbit (p. 194) now as a rabbit, now as a duck; we can see the Necker cube, now with one face forward, now with another; we can see a cube drawing (p. 193) as a box, a wire frame, etc. How, if at all, does our visual experience change? The experience is much more elusive than is anything like the feeling of a headache, the hearing of a sound, the visual experience of a blue patch. The corresponding 'aspects' of meaning would seem to be introspectively even more elusive.

Similarly, although some of the passages in §§156–78 seem to debunk the idea of a conscious special experience of 'being guided' (when reading) altogether, it seems wrong to think of it as totally dismissed. For example, in §160, Wittgenstein speaks both of the 'sensation of saying something he has learnt by heart' and of the 'sensation of reading', though the point of the paragraph is that the presence or absence of such sensations is not what constitutes the distinction between reading, saying something by heart, and yet something else. To some extent, I think Wittgenstein's discussion may have a certain ambivalence. Nevertheless, some relevant points made are these: (i) Whatever an 'experience of being guided' (in reading) may be, it is not something with a gross and introspectible qualitative character, like a headache (contrary to Hume). (ii) In particular cases of reading, we may feel definite and introspectible experiences, but these are different and distinct experiences, peculiar to each individual case, not a single experience present in all cases. (In the same way, Wittgenstein speaks of various introspectible 'mental processes' that *in particular circumstances* occur when I understand a word – see §§151–5, but more of these *is* the 'process' of understanding, indeed understanding is not a 'mental process' – see pp. 49–51 below. The discussion of reading, which follows §§151–5 immediately, is meant to illustrate these points. (iii) Perhaps most important, whatever the elusive feeling of being guided may be, its presence or absence is not constitutive of whether I am reading or not. See, for example, the cases mentioned above in the text, of the pupil learning to read and of the person under the influence of a drug.

Rush Rhees, in his preface to *The Blue and Brown Books* (Basil Blackwell, Oxford and Harper and Brothers, New York, 1958, xiv+185 pp.) emphasizes (see pp. xii–xiv) the problem created for Wittgenstein by 'meaning blindness', and he emphasizes that the discussion of 'seeing something as something' in section xi of the second

can be brief because this particular Wittgensteinian lesson has been relatively well learned, perhaps too well learned. But some points should be noted. First, to repeat, the method of the investigation, and of the thought-experiments is deeply introspective: it is exactly the kind of investigation a strict psychological behaviorist would *prohibit*.[30] Second, although Wittgenstein does conclude that behavior, and dispositions to behavior, lead us to *say* of a person that he is reading, or adding, or whatever, this should not, in my opinion, be misconstrued as an endorsement of the dispositional theory: he does not say that reading or adding *is* a certain disposition to behavior.[31]

---

part of *Philosophical Investigations* is motivated by an attempt to deal with the elusive question. Earlier portions of the *Investigations* repudiate traditional pictures of internal, qualitative states of meaning and understanding; but later Wittgenstein seems, as Rhees says, to be worried that he may be in danger of replacing the classical picture by an overly mechanistic one, though certainly he still repudiates any idea that a certain qualitative experience *is* what constitutes my using words with a certain meaning. Could there be a 'meaning blind' person who operated with words just as we do? If so, would we say that he is as much in command of the language as we? The official answer, as given in our main text, is 'yes'; but perhaps the answer should be, "Say what you want, as long as you know the facts." It is not clear that the problem is entirely resolved. Note that here, too, the discussion is introspective, based on an investigation of our own phenomenal experience. It is not the kind of investigation that would be undertaken by a behaviorist. No doubt the matter deserves a careful and extended treatment.

[30] §314 says: "It shows a fundamental misunderstanding, if I am inclined to study the headache I have now in order to get clear about the fundamental philosophical problem of sensation." If this remark is to be consistent with Wittgenstein's frequent practice as outlined in the text above and note 29, it *cannot* be read as *generally* condemning the philosophical use of introspective reflections on the phenomenology of our experience.

[31] I should not deny that Wittgenstein has important affinities to behaviorism (as to finitism – see pp. 196–7 below). Such a famous slogan as "My attitude toward him is an attitude towards a soul (*Seele*). I am not of the *opinion* that he has a soul" (p. 178) sounds much too behavioristic for me. I personally would like to think that anyone who does not think of me as conscious is wrong about the facts, not simply 'unfortunate', or 'evil', or

Wittgenstein's conviction of the contrast between states of understanding, reading and the like, and 'genuine', introspectible mental states or processes is so strong that it leads him – who is often regarded as a (or the) father of 'ordinary language philosophy', and who emphasizes the importance of respect for the way language is actually used – into some curious remarks about ordinary usage. Consider §154: "In the sense in which there are processes (including mental processes) which are characteristic of understanding, understanding is not a mental process. (A pain's growing more and less; the hearing of a tune or sentence: these are mental processes.)" Or again, at the bottom of p. 59, "'Understanding a word': a state. But a *mental* state? – Depression, excitement, pain, are called mental states. Carry out a grammatical investigation . . ." The terms 'mental state' and 'mental process' have a somewhat theoretical flavor, and I am not sure how firmly one can speak of their 'ordinary' use. However, my own linguistic intuitions do not entirely agree with Wittgenstein's remarks.[32] Coming to understand, or learning, seems to me to

---

even 'monstrous' or 'inhuman', in his 'attitude' (whatever that might mean).

(If '*Seele*' is translated as 'soul', it might be thought that the 'attitude' (*Einstellung*') to which Wittgenstein refers has special religious connotations, or associations with Greek metaphysics and the accompanying philosophical tradition. But it is clear from the entire passage that the issue relates simply to the difference between my 'attitude' toward a conscious being, and toward an automaton, even though one of the paragraphs refers specifically to the religious doctrine of the immortality of the soul ('*Seele*'). Perhaps in some respects 'mind' might be a less misleading translation of '*Seele*' in the sentence quoted above, since for the contemporary English speaking philosophical reader it is somewhat less loaded with special philosophical and religious connotations. I feel that this may be so even if 'soul' captures the flavor of the German '*Seele*' better than 'mind'. Anscombe translates '*Seele*' and its derivatives sometimes as 'soul', sometimes as 'mind', depending on context. The problem really seems to be that German has only '*Seele*' and '*Geist*' to do duty where an English speaking philosopher would use 'mind'. See also the postscript below, note 11.

[32] These are my intuitions in English. I have no idea whether any differences

be a 'mental process' if anything is. A pain's growing more and less, and especially the hearing of a tune or sentence, are probably not ordinarily thought of as 'mental' processes at all. Although depression and anxiety would ordinarily be called 'mental' states, pain (if genuine physical pain is meant) is probably *not* a 'mental' state. ("It's all in your mind" means that no genuine physical pain is present.) But Wittgenstein's concern is not really with usage but with a philosophical terminology. 'Mental states' and 'mental processes' are those introspectible 'inner' contents that I can find in my mind, or that God could find if he looked into my mind.[33] Such

with the German ('*seelischer Vorgang*' and '*seelischer Zustand*'), in nuance or usage, affect the matter.

[33] Or so it would seem from the passages quoted. But the denial that understanding is a 'mental process' in §154 is preceded by the weaker remark, "Try not to think of understanding as a 'mental process' at all – for that is the expression that confuses you. In itself, this seems to say that thinking of understanding as a 'mental process' leads to misleading philosophical pictures, but not necessarily that it is wrong. See also §§305–6 "But you surely cannot deny that, for example, in remembering, an inner process takes place." What gives the impression that we want to deny anything? . . . What we deny is that the picture of the inner process gives us the correct use of the word 'to remember' . . . Why should I deny that there is a mental process? But "There has just taken place in me the mental process of remembering . . ." means nothing more than: "I have just remembered . . ." To deny the mental process would mean to deny the remembering; to deny that anyone ever remembers anything.' *This* passage gives the impression that *of course* remembering is a 'mental process' if anything is, but that this ordinary terminology is philosophically misleading. (The German here is '*geistiger Vorgang*' while in the earlier passages it was '*seelischer Vorgang*' (§154) and '*seelischer Zustand*' (p. 59), but as far as I can see, this has no significance beyond stylistic variation. It is possible that the fact that Wittgenstein speaks here of remembering, while earlier he had spoken of understanding is significant, but even this seems to me to be unlikely. Note that in §154, the genuine 'mental processes' are a pain's growing more or less, the hearing of a tune or sentence – processes with an 'introspectible quality' in the sense we have used the phrase. For Wittgenstein remembering is not a process like these, even though, as in the case of understanding in §154, there may be processes with introspectible qualities that take place when

phenomena, inasmuch as they are introspectible, 'qualitative' states of the mind, are not subject to immediate sceptical challenge of the present type. Understanding is not one of these.

Of course the falsity of the 'unique introspectible state' view of meaning plus must have been implicit from the start of the problem. If there really were an introspectible state, like a headache, of meaning addition by 'plus' (and if it really could have the justificatory role such a state ought to have), it would have stared one in the face and would have robbed the sceptic's challenge of any appeal. But given the force of this challenge, the need philosophers have felt to posit such a state and the loss we incur when we are robbed of it should be apparent. Perhaps we may try to recoup, by arguing that meaning addition by 'plus' is a state even more *sui generis* than we have argued before. Perhaps it is simply a primitive state, not to be assimilated to sensations or headaches or any 'qualitative' states, nor to be assimilated to dispositions, but a state of a unique kind of its own.

Such a move may in a sense be irrefutable, and taken in an appropriate way Wittgenstein may even accept it. But it seems desperate: it leaves the nature of this postulated primitive state – the primitive state of 'meaning addition by "plus"' – completely mysterious. It is not supposed to be an introspectible state, yet we supposedly are aware of it with some fair degree of certainty whenever it occurs. For how else can each of us be confident that he *does*, at present, mean addition by 'plus'? Even more important is the logical difficulty implicit in Wittgenstein's sceptical argument. I think that Wittgenstein argues, not merely as we have said hitherto, that introspection shows that the alleged 'qualitative' state of understanding is a

we remember. Assuming that the examples given in §154 are meant to be typical 'mental processes', the examples would be very misleading unless remembering were taken not to be a 'mental process' in the sense of §154. Remembering, like understanding, is an 'intentional' state (see note 19 above) subject to Wittgenstein's sceptical problem.) (See also the discussion of 'incorporeal processes' in §339.)

chimera, but also that it is logically impossible (or at least that there is a considerable logical difficulty) for there to be a state of 'meaning addition by "plus"' at all.

Such a state would have to be a finite object, contained in our finite minds.[34] It does not consist in my explicitly thinking of each case of the addition table, nor even of my encoding each separate case in the brain: we lack the capacity for that. Yet (§195) "in a *queer* way" each such case already is "in some sense present". (Before we hear Wittgenstein's sceptical argument, we surely suppose – unreflectively – that something like this is indeed the case. Even now I have a strong inclination to think this somehow must be right.) What can that sense be? Can we conceive of a finite state which *could* not be interpreted in a quus-like way? How could that be? The proposal I am now discussing brushes such questions under the rug, since the nature of the supposed 'state' is left

[34] We have stressed that I think of only finitely many cases of the addition table. Anyone who claims to have thought of infinitely many cases of the table is a liar. (Some philosophers – probably Wittgenstein – go so far as to say that they see a conceptual incoherence in the supposition that anyone thought of infinitely many such cases. We need not discuss the merits of this strong view here as long as we acknowledge the weaker claim that as a matter of fact each of us thinks of only finitely many cases.) It is worth noting, however, that although it is useful, following Wittgenstein himself, to *begin* the presentation of the puzzle with the observation that I have thought of only finitely many cases, it appears that in principle this particular ladder can be kicked away. Suppose that I had explicitly thought of *all* cases of the addition table. How can this help me answer the question '68+57'? Well, looking back over my own mental records, I find that I gave myself explicit directions. "If you are ever asked about '68+57', reply '125'!" Can't the sceptic say that these directions, too, are to be interpreted in a non-standard way? (See *Remarks on the Foundations of Mathematics*, I, §3: "If I know it *in advance*, what use is this knowledge to me later on? I mean: how do I know what to do with this earlier knowledge when the step is actually taken?") It would appear that, if finiteness is relevant, it comes more crucially in the fact that "justifications must come to an end somewhere" than in the fact that I think of only finitely many case of the addition table, even though Wittgenstein stresses both facts. Either fact can be used to develop the sceptical paradox; both are important.

mysterious. "But" – to quote the protest in §195 more fully – "I don't mean that what I do now (in grasping a sense) determines the future use *causally* and as a matter of experience, but that in a *queer* way, the use itself is in some sense present." A causal determination is the kind of analysis supposed by the dispositional theorist, and we have already seen that that is to be rejected. Presumably the relation now in question grounds some entailment roughly like: "If I now mean addition by 'plus'; then, if I remember this meaning in the future and wish to accord with what I meant, and do not miscalculate, then when asked for '68+57', I will respond '125'." If Hume is right, of course, no past state of my mind can entail that I will give any particular response in the future. But that I meant 125 in the past does not itself entail this; I must remember what I meant, and so on. Nevertheless it remains mysterious exactly how the existence of *any* finite past state of my mind could entail that, if I wish to accord with it, and remember the state, and do not miscalculate, I must give a determinate answer to an arbitrarily large addition problem.[35]

Mathematical realists, or 'Platonists', have emphasized the non-mental nature of mathematical entities. The addition function is not in any particular mind, nor is it the common property of all minds. It has an independent, 'objective', existence. There is then no problem – as far as the present considerations go – as to how the addition function (taken, say, as a set of triples)[36] contains within it all its instances, such as the triple (68, 57, 125). This simply is in the nature of the mathematical object in question, and it may well be an infinite

[35] See p. 218: "Meaning it is not a process which accompanies a word. For no process could have the consequences of meaning." This aphorism makes the general point sketched in the text. No process can entail what meaning entails. In particular, no process could entail the rough conditional stated above. See the discussion below, pp. 93–4, of Wittgenstein's view of these conditionals.

[36] Of course Frege would not accept the identification of a function with a set of triples. Such an identification violates his conception of functions as 'unsaturated'. Although this complication is very important for Frege's philosophy, it can be ignored for the purposes of the present presentation.

object. The proof that the addition function contains such a triple as (68, 57, 125) belongs to mathematics and has nothing to do with meaning or intention.

Frege's analysis of the usage of the plus sign by an individual posits the following four elements: (a) the addition function, an 'objective' mathematical entity; (b) the addition sign '+', a linguistic entity; (c) the 'sense' of this sign, an 'objective' abstract entity like the function; (d) an idea in the individual's mind associated with the sign. The idea is a 'subjective' mental entity, private to each individual and different in different minds. The 'sense', in contrast, is the same for all individuals who use '+' in the standard way. Each such individual grasps this sense by virtue of having an appropriate idea in his mind. The 'sense' in turn *determines* the addition function as the *referent* of the '+' sign.

There is again no special problem, for this position, as to the relation between the sense and the referent it determines. It simply is in the nature of a sense to determine a referent. But ultimately the sceptical problem cannot be evaded, and it arises precisely in the question how the existence in my mind of any mental entity or idea can *constitute* 'grasping' any particular sense rather than another. The idea in my mind is a finite object: can it not be interpreted as determining a quus function, rather than a plus function? Of course there may be another idea in my mind, which is supposed to constitute its act of *assigning* a particular interpretation to the first idea; but then the problem obviously arises again at this new level. (A rule for interpreting a rule again.) And so on. For Wittgenstein, Platonism is largely an unhelpful evasion of the problem of how our finite minds can give rules that are supposed to apply to an infinity of cases. Platonic objects may be self-interpreting, or rather, they may need no interpretation; but ultimately there must be some mental entity involved that raises the sceptical problem. (This brief discussion of Platonism is meant for those interested in the issue. If it is so brief that you find it obscure, ignore it.)

3

The Solution and the 'Private Language' Argument

The sceptical argument, then, remains unanswered. There can be no such thing as meaning anything by any word. Each new application we make is a leap in the dark; any present intention could be interpreted so as to accord with anything we may choose to do. So there can be neither accord, nor conflict. This is what Wittgenstein said in §202.

Wittgenstein's sceptical problem is related to some work of two other recent writers who show little direct influence from Wittgenstein. Both have already been mentioned above. The first is W. V. Quine,[37] whose well-known theses of the indeterminacy of translation and the inscrutability of reference also question whether there are any objective facts as to what we mean. If I may anticipate matters that the present exposition has not yet introduced, Quine's emphasis on agreement is obviously congenial to Wittgenstein's view.[38] So

[37] See pp. 14–15 above, and note 10.

[38] For 'agreement' and the related notion of 'form of life' in Wittgenstein, see pp. 96–8 below. Quine, *Word and Object*, p. 27, characterizes language as "the complex of present dispositions to verbal behavior, in which speakers of the same language have perforce come to resemble one another"; also, see §2, *Word and Object*, pp. 5–8. Some of the major

is his rejection of any notion that inner 'ideas' or 'meanings' guide our linguistic behavior. However, there are differences. As I have remarked above, Quine bases his argument from the outset on behavioristic premises. He would never emphasize introspective thought experiments in the way Wittgenstein does, and he does not think of views that posit a private inner world as in need of elaborate refutation. For Quine, the untenability of any such views should be obvious to anyone who accepts a modern scientific outlook. Further, since Quine sees the philosophy of language within a hypothetical framework of behavioristic psychology, he thinks of problems about meaning as problems of disposition to behavior. This orientation seems to have consequences for the form of Quine's problem as opposed to Wittgenstein's. The important problem for Wittgenstein is that my present mental state does not appear to determine what I *ought* to do in the future. Although I may *feel* (now) that something in my head corresponding to the word 'plus' mandates a determinate response to any new pair of arguments, in fact nothing in my head does so. Alluding to one of Wittgenstein's earliest examples, 'ostensive' learning of the color word 'sepia' (§§28–30),[39] Quine protests against Wittgenstein that, given our 'inborn propensity to find one stimulation qualitatively more akin to a second stimulation than to a third' and sufficient conditioning 'to eliminate wrong generalizations', eventually the term will be learnt: ". . . in principle nothing more is needed in learning 'sepia' than in any conditioning or induction."[40] By "learning 'sepia'", Quine means developing the right disposition to apply 'sepia' in particular cases. It should be clear from Wittgenstein's text that he too is aware, indeed emphasizes, that in practice there need be no difficulty

concepts of *Word and Object*, such as that of 'observation sentence', depend on this uniformity in the community. Nevertheless, agreement seems to have a more crucial role in Wittgenstein's philosophy than in Quine's.

[39] This example is discussed below. See pp. 83–4 and note 72.

[40] Quine, *Ontological Relativity and Other Essays*, p. 31.

in this sense about the learning of 'sepia'. The fundamental problem, as I have stated it earlier, is different: whether my actual dispositions are 'right' or not, is there anything that mandates what they *ought* to be? Since Quine formulates the issues dispositionally, this problem cannot be stated within his framework. For Quine, since any fact as to whether I mean plus or quus will show up in my behavior, there is no question, given my disposition, as to what I mean.

It has already been argued above that such a formulation of the issues seems inadequate. My actual dispositions are not infallible, nor do they cover all of the infinitely many cases of the addition table. However, since Quine does see the issues in terms of dispositions, he is concerned to show that even if dispositions were ideally seen as infallible and covering all cases, there are still questions of interpretation that are left undetermined. First, he argues roughly that the interpretation of sufficiently 'theoretical' utterances, not direct observation reports, is undetermined even by all my ideal dispositions. Further, he seeks to show by examples such as 'rabbit' and 'rabbit-stage' that, even given fixed interpretation of our sentences as wholes and certainly given all our ideal dispositions to behavior, the interpretation (reference) of various lexical items is still not fixed.[41] These are interesting claims, distinct from Wittgenstein's. For those of us who are not as behavioristically inclined as Quine, Wittgenstein's problem may lead to a new look at Quine's theses. Given Quine's own formulation of his theses, it appears open to a non-behaviorist to regard his arguments, *if* he accepts them, as demonstrations that any behavioristic account of meaning must be inadequate – it cannot even distinguish between a word meaning rabbit and one meaning rabbit-stage. But if Wittgenstein is right, and no amount of access to my mind can reveal whether I mean plus or quus, may the same not hold for rabbit and rabbit-stage? So perhaps Quine's problem arises even for non-behaviorists. This is not the place to explore the matter.

[41] Roughly, the first assertion is the 'indeterminacy of translation', while the second is the 'inscrutability of reference'.

Nelson Goodman's discussion of the 'new riddle of induction' also deserves comparison with Wittgenstein's work.[42] Indeed, although Quine, like Wittgenstein, and unlike Goodman in his treatment of the 'new riddle', directly concerns himself with a sceptical doubt about meaning, the basic strategy of Goodman's treatment of the 'new riddle' is strikingly close to Wittgenstein's sceptical arguments. In this respect, his discussion is much closer to Wittgenstein's scepticism than is Quine's treatment of 'indeterminacy'. Although our paradigm of Wittgenstein's problem was formulated for a mathematical problem, it was emphasized that it is completely general and can be applied to any rule or word. In particular, if it were formulated for the language of color impressions, as Wittgenstein himself suggests, Goodman's 'grue' or something similar, would play the role of 'quus'.[43] But the problem would not be Goodman's about induction – "Why not predict that grass, which has been grue in the past, will be grue in the future?" – but Wittgenstein's about meaning: "Who is to say that in the past I didn't mean grue by 'green', so that now I should call the sky, not the grass, 'green'?" Although Goodman concentrates on the problem about induction and largely ignores the problem about meaning,[44] his discussions are occasionally suggestive for

[42] See Goodman, *Fact, Fiction, and Forecast*, p. 13b, n. 1. See also the papers in part VII ("Induction") in *Problems and Projects* (Bobbs-Merrill, Indianapolis and New York, 1972, xii+463 pp.)

[43] For 'grue', see page 20 and footnotes 14 and 15 above. My memory about my own thought processes years ago is weak, but it seems likely that I may have been inspired to formulate Wittgenstein's problem in terms of 'quus' by Goodman's analogous use of 'grue'. I do remember that, at the time I first thought about the problem, I was struck by the analogy between Wittgenstein's discussion and Goodman's (as others have been as well).

[44] In part Goodman's discussion of the problem seems to presuppose that the extension of each predicate ('green', 'grue'), etc., is known, and that this question does not itself get entangled in the 'new riddle of induction'. Sydney Shoemaker, "On Projecting the Unprojectible," *The Philosophical Review*, vol. 84 (1975), pp. 178–219, questions whether such a separation is possible (see his concluding paragraph). I have not yet made a careful study of Shoemaker's argument.

Wittgenstein's problem as well.[45] In fact, I personally suspect that serious consideration of Goodman's problem, as he formulates it, may prove impossible without consideration of Wittgenstein's.[46]

[45] See his "Positionality and Pictures," *The Philosophical Review*, vol. 69 (1960), pp. 523–5, reprinted in *Problems and Projects*, pp. 402–4. See also Ullian, "More on 'Grue' and Grue," and *Problems and Projects*, pp. 408–9 (comments on Judith Thompson).

"Seven Strictures on Similarity," *Problems and Projects*, pp. 437–46, has in places a Wittgensteinian flavor. For Goodman, as for Wittgenstein, what we call 'similar' (for Wittgenstein: even 'the same') is exhibited in our own practice and cannot explain it. (For an exposition of Wittgenstein's position, see section 3 below.)

One issue arises here. Does Wittgenstein's position depend on a denial of 'absolute similarity'? To the extent that we use 'similarity' simply to *endorse* the way we actually go on, it does. But it is important to see that, even if 'absolutely similar' had a fixed meaning in English, and 'similar' did not need to be filled in by a specification of the 'respects' in which things are similar, the sceptical problem would not be solved. When I learn 'plus', I could not simply give myself some finite number of examples and continue: 'Act similarly when confronted with any addition problem in the future.' Suppose that, on the ordinary meaning of 'similar' this construction is completely determinate, and that one does not hold the doctrine that various alternative ways of acting can be called 'similar', depending on how 'similar' is filled out by speaking of a respect in which one or another way of acting can be called 'similar' to what I did before. Even so, the sceptic can argue that by 'similar' I meant *quimilar*, where two actions are quimilar if . . . See also the discussion of 'relative identity', note 13 above.

[46] Briefly: Goodman insists that there is no sense that does not beg the question according to which 'grue' is 'temporal', or 'positional', and 'green' is not; if either of the pairs 'blue-green' and 'grue-bleen' is taken as primitive, the predicates of the other pair are 'temporally' definable in terms of it (see *Fact, Fiction, and Forecast*, pp. 77–80). Nevertheless, intuitively it does seem clear that 'grue' is positional in a sense that 'green' is not. Perhaps that sense can be brought out by the fact that 'green', but not 'grue', is learned (learnable?) ostensively by a sufficient number of samples, without reference to time. It would seem that a reply to this argument should take the form. "Who is to say that it is not 'grue' that others (or even, myself in the past?) learned by such ostensive training?" But this leads directly to Wittgenstein's problem. The papers cited in the previous footnote are relevant. (It is true, however, that problems like Goodman's can arise for competing predicates that do not appear, even intuitively, to be defined positionally.)

Wittgenstein has invented a new form of scepticism. Personally I am inclined to regard it as the most radical and original sceptical problem that philosophy has seen to date, one that only a highly unusual cast of mind could have produced. Of course he does not wish to leave us with his problem, but to solve it: the sceptical conclusion is insane and intolerable. It is his solution, I will argue, that contains the argument against 'private language'; for allegedly, the solution will not admit such a language. But it is important to see that his achievement in posing this problem stands on its own, independently of the value of his own solution of it and the resultant argument against private language. For, if we see Wittgenstein's problem as a real one, it is clear that he has often been read from the wrong perspective. Readers, my previous self certainly included, have often been inclined to wonder: "How can he prove private language impossible? How can I possibly have any difficulty identifying my own sensations? And if there were a difficulty, how could 'public' criteria help me? I must be in pretty bad shape if I needed external *help* to identify my own sensations!"[47] But if I am right, a proper

[47] Especially for those who know some of the literature on the 'private language argument', an elaboration of this point may be useful. Much of this literature, basing itself on Wittgenstein's discussions following §243, thinks that without some external check on my identification of my own sensations, I would have no way of knowing that I have identified a given sensation correctly (in accord with my previous intentions). (The question has been interpreted to be, "How do I know I am right that this is pain?", or it might be, "How do I know that I am applying the right rule, using 'pain' as I had intended it"? See note 21 above.) But, it is argued, if I have no way of knowing (on one of these interpretations) whether I am making the right identification, it is meaningless to speak of an identification at all. To the extent that I rely on my own impressions or memories of what I meant by various sensation signs for support, I have no way of quelling these doubts. Only others, who recognize the correctness of my identification through my external behavior, can provide an appropriate external check.

A great deal could be said about the argument just obscurely summarized, which is not easy to follow even on the basis of longer presentations in the literature. But here I wish to mention one reaction: If

I really were in doubt as to whether I could identify any sensations correctly, how would a connection of my sensations with external behavior, or confirmation by others, be of any help? Surely I can identify that the relevant external behavior has taken place, or that others are confirming that I do indeed have the sensation in question, only because I can identify relevant sensory impressions (of the behavior, or of others confirming that I have identified the sensation correctly). My ability to make any identification of any external phenomenon rests on my ability to identify relevant sensory (especially visual) impressions. If I were to entertain a *general* doubt of my ability to identify any of my own mental states, it would be impossible to escape from it.

It is in this sense that it may appear that the argument against private language supposes that I need external *help* to identify my own sensations. For many presentations of the argument make it appear to depend on such a general doubt of the correctness of all my identifications of inner states. It is argued that since any identification I make needs some kind of verification for correctness, a verification of one identification of an inner state by another such identification simply raises the very same question (whether I am making a correct identification of my sensations) over again. As A. J. Ayer, in his well know exchange with Rush Rhees ("Can there be a Private Language?" *Proceedings of the Aristotelian Society*, supp. vol. 28 (1954), pp. 63–94, reprinted in Pitcher (ed.), *Wittgenstein: The Philosophical Investigations*, pp. 251–85, see especially p. 256), summarizes the argument, "His claim to recognize the object [the sensation], his belief that it really is the same, is not to be accepted unless it can be backed by further evidence. Apparently, too, this evidence must be public . . . Merely to check one private sensation by another would not be enough. For if one cannot be trusted to recognize one of them, neither can one be trusted to recognize the other." The argument concludes that I can make a genuine verification of the correctness of my identification only if I break out of the circle of 'private checks' to some publicly accessible evidence. But if I were so sceptical as to doubt *all* my identifications of inner states , how could anything public be of any help? Does not my recognition of anything public depend on the recognition of my inner states? As Ayer puts it (immediately following the earlier quotation), "But unless there is some thing that one is allowed to recognize, no test can ever be completed . . . I check my memory of the time at which the train is due to leave by visualizing a page of the time-table; and I am required to check this in its turn by looking up the page. [He is alluding to §265.] But unless I can trust my eyesight at this point, unless I can recognize the figures that I see written down, I am still no better off . . . Let the object to which I am attempting to refer be as public as you please . . . my assurance that I am using the word correctly . . . must in the end rest on the testimony of the senses. It is through

orientation would be the opposite. The main problem is *not*, "How can we show private language – or some other special form of language – to be *impossible?*"; rather it is, "How can we show *any language* at all (public, private, or what-have-you) to be *possible?*"[48] It is not that calling a sensation 'pain' is easy, and Wittgenstein must invent a difficulty.[49] On the contrary, Wittgenstein's main problem is that it appears that he has shown *all* language, *all* concept formation, to be impossible, indeed unintelligible.

It is important and illuminating to compare Wittgenstein's new form of scepticism with the classical scepticism of Hume; there are important analogies between the two. Both develop a sceptical paradox, based on questioning a certain *nexus* from past to future. Wittgenstein questions the nexus between past 'intention' or 'meanings' and present practice: for example, between my past 'intentions' with regard to 'plus' and my present computation '68+57=125'. Hume questions two other nexuses, related to each other: the causal nexus whereby a past event necessitates a future one, and the inductive inferential nexus from the past to the future.

hearing what other people say, or through seeing what they write, or observing their movements, that I am enabled to conclude that their use of the word agrees with mine. But if without further ado I can recognize such noises or shapes or movements, why can I not also recognize a private sensation?"

Granted that the private language argument is presented simply in this form, the objection seems cogent. Certainly it once seemed to me on some basis such as this that the argument against private language *could* not be right. Traditional views, which are very plausible unless they are decisively rebutted, hold that all identifications rest on the identification of sensations. The sceptical interpretation of the argument in this essay, which does not allow the notion of an identification to be taken for granted, makes the issue very different. See the discussion, on pp. 67–8 below, of an analogous objection to Hume's analysis of causation.

[48] So put, the problem has an obvious Kantian flavor.

[49] See especially the discussions of 'green' and 'grue' above, which plainly could carry over to pain (let 'pickle' apply to pains before *t*, and tickles thereafter!); but it is clear enough by now that the problem is completely general.

The analogy is obvious. It has been obscured for several reasons. First, the Humean and the Wittgensteinian problems are of course distinct and independent, though analogous. Second, Wittgenstein shows little interest in or sympathy with Hume: he has been quoted as saying that he could not read Hume because he found it "a torture".[50] Furthermore, Hume is the prime source of some ideas on the nature of mental states that Wittgenstein is most concerned to attack.[51] Finally (and probably most important), Wittgenstein never avows, and almost surely would not avow, the label 'sceptic', as Hume explicitly did. Indeed, he has often appeared to be a 'common-sense' philosopher, anxious to defend our ordinary conceptions and dissolve traditional philosophical doubts. Is it not Wittgenstein who held that philosophy only states what everyone admits?

Yet even here the difference between Wittgenstein and Hume should not be exaggerated. Even Hume has an important strain, dominant in some of his moods, that the philosopher never questions ordinary beliefs. Asked whether he "be really one of those sceptics, who hold that all is uncertain", Hume replies "that this question is entirely superfluous, and that neither I, nor any other person, was ever sincerely and constantly of that opinion".[52] Even more forcefully, discussing the problem of the external world: "We

[50] Karl Britton, "Portrait of a Philosopher," *The Listener*, LIII, no. 1372 (June 16, 1955), p. 1072, quoted by George Pitcher, *The Philosophy of Wittgenstein* (Prentice Hall, Englewood Cliffs, NJ, 1964, viii+340 pp), p. 325.

[51] Much of Wittgenstein's argument can be regarded as an attack on characteristically Humean (or classical empiricist) ideas. Hume posits an introspectible qualitative state for each of our psychological states (an 'impression'). Further, he thinks that an appropriate 'impression' or 'image' can constitute an 'idea', without realizing that an image in no way tells us how it is to be applied. (See the discussion of determining the meaning of 'green' with an image on p. 20 above and the corresponding discussion of the cube on pp. 42–3 above.) Of course the Wittgensteinian paradox is, among other things, a strong protest against such suppositions.

[52] David Hume, *A Treatise of Human Nature* (ed. L. A. Selby-Bigge,

may well ask, *What causes induce us to believe in the existence of body?* but 'tis in vain to ask, *Whether there be body or not?* That is a point, which we must take for granted in all our reasonings."[53] Yet this oath of fealty to common sense begins a section that otherwise looks like an argument that the common conception of material objects is irreparably incoherent!

When Hume is in a mood to respect his professed determination never to deny or doubt our common beliefs, in what does his 'scepticism' consist? First, in a sceptical *account* of the causes of these beliefs; and second, in sceptical analyses of our common notions. In some ways Berkeley, who did not regard his own views as sceptical, may offer an even better analogy to Wittgenstein. At first blush, Berkeley, with his denial of matter, and of any objects 'outside the mind' seems to be *denying* our common beliefs; and for many of us the impression persists through later blushes. But not for Berkeley. For him, the impression that the common man is committed to matter and to objects outside the mind derives from an erroneous metaphysical interpretation of common talk. When the common man speaks of an 'external material object' he does not really mean (as we might say *sotto voce*) an *external material object* but rather he means something like 'an idea produced in me independently of my will'.[54]

Clarendon Press, Oxford, 1888), Book I, Part IV, Section I (p. 183 in the Selby-Bigge edition).

[53] Hume, *ibid.*, Book I, Part II, Section II (p. 187 in the Selby-Bigge edition). Hume's occasional affinities to 'ordinary language'-philosophy should not be overlooked. Consider the following: "Those philosophers, who have divided human reason into *knowledge and probability*, and have defined the first to be *that evidence, which arises from a comparison of ideas*, are obliged to comprehend all our arguments from causes or effects under the general term of probability. But tho' everyone be free to use his terms in what sense he pleases . . . 'tis however certain, that in common discourse we readily affirm, that many arguments from causation exceed probability, and may be received as a superior kind of evidence. One would appear ridiculous, who would say, that 'tis only probable the sun will rise tomorrow, or that all men must dye . . ." (*ibid.*, Book I, Part III, Section XI, p. 124 in the Selby-Bigge edition).

[54] George Berkeley, *The Principles of Human Knowledge*, §§29–34. Of course

Berkeley's stance is not uncommon in philosophy. The philosopher advocates a view apparently in patent contradiction to common sense. Rather than repudiating common sense, he asserts that the conflict comes from a philosophical misinterpretation of common language – sometimes he adds that the misinterpretation is encouraged by the 'superficial form' of ordinary speech. He offers his own analysis of the relevant common assertions, one that shows that they do not really say what they seem to say. For Berkeley this philosophical strategy is central to his work. To the extent that Hume claims that he merely analyses common sense and does not oppose it, he invokes the same strategy as well. The practice can hardly be said to have ceased today.[55]

Personally I think such philosophical claims are almost invariably suspect. What the claimant calls a 'misleading philosophical misconstrual' of the ordinary statement is probably the natural and correct understanding. The real misconstrual comes when the claimant continues, "All the ordinary man really means is . . ." and gives a sophisticated analysis compatible with his own philosophy. Be this as it may, the important point for present purposes is that Wittgenstein makes a Berkeleyan claim of this kind. For – as we shall see – his solution to his own sceptical problem begins by agreeing with the sceptics that there is no 'superlative fact' (§192) about my mind that constitutes my meaning addition by 'plus' and determines in advance what I should do to accord with this meaning. But, he claims (in §§183–93), the appearance that our ordinary concept of meaning demands such a fact is based on a philosophical misconstrual – albeit a natural one –

the characterization may be oversimplified, but it suffices for present purposes.

[55] It is almost 'analytic' that I cannot produce a common contemporary example that would not meet with vigorous opposition. Those who hold the cited view will argue that, in this case, their analyses of ordinary usage are really correct. I have no desire to enter into an irrelevant controversy here, but I myself find that many of the 'topic-neutral' analyses of discourse about the mind proposed by contemporary materialists are just the other side of the Berkeleyan coin.

of such ordinary expressions as 'he meant such-and-such', 'the steps are determined by the formula', and the like. How Wittgenstein construes these expressions we shall see presently. For the moment let us only remark that Wittgenstein thinks that any construal that looks for something in my present mental state to differentiate between my meaning addition or quaddition, or that will consequently show that in the future I should say '125' when asked about '68+57', *is* a misconstrual and attributes to the ordinary man a notion of meaning that *is* refuted by the sceptical argument. "We are," he says in §194 – note that Berkeley could have said just the same thing! – "like savages, primitive people, who hear the expressions of civilized men, put a false interpretation on them, and then draw the queerest conclusions from it." Maybe so. Personally I can only report that, in spite of Wittgenstein's assurances, the 'primitive' interpretation often sounds rather good to me . . .

In his *Enquiry*, after he has developed his "Sceptical Doubts Concerning the Operations of the Understanding", Hume gives his "Sceptical Solution of These Doubts". What is a 'sceptical' solution? Call a proposed solution to a sceptical philosophical problem a *straight* solution if it shows that on closer examination the scepticism proves to be unwarranted; an elusive or complex argument proves the thesis the sceptic doubted. Descartes gave a 'straight' solution in this sense to his own philosophical doubts. An *a priori* justification of inductive reasoning, and an analysis of the causal relation as a genuine necessary connection or nexus between pairs of events, would be straight solutions of Hume's problems of induction and causation, respectively. A *sceptical* solution of a sceptical philosophical problem begins on the contrary by conceding that the sceptic's negative assertions are unanswerable. Nevertheless our ordinary practice or belief is justified because – contrary appearances notwithstanding – it need not require the justification the sceptic has shown to be untenable. And much of the value of the sceptical argument consists precisely in the fact that he has shown that an ordinary practice, if it is to be

defended at all, cannot be defended in a certain way. A sceptical solution may also involve – in the manner suggested above – a sceptical analysis or account of ordinary beliefs to rebut their *prima facie* reference to a metaphysical absurdity.

The rough outlines of Hume's sceptical solution to his problem are well known.[56] Not an *a priori* argument, but custom, is the source of our inductive inferences. If *A* and *B* are two types of events which we have seen constantly conjoined, then we are conditioned – Hume is a grandfather of this modern psychological notion – to expect an event of type *B* on being presented with one of type *A*. To say of a particular event *a* that it caused another event *b* is to place these two events under two types, *A* and *B*, which we expect to be constantly conjoined in the future as they were in the past. The idea of necessary connection comes from the 'feeling of customary transition' between our ideas of these event types.

The philosophical merits of the Humean solution are not our present concern. Our purpose is to use the analogy with the Humean solution to illuminate Wittgenstein's solution to his own problem. For comparative purposes one further consequence of Hume's sceptical solution should be noted. Naively, one might suppose that whether a particular event *a* causes another particular event *b*, is an issue solely involving the events *a* and *b* alone (and their relations), and involves no other events: If Hume is right, this is not so. Even if God were to look at the events, he would discern nothing relating them other than that one succeeds the other. Only when the particular events *a* and *b* are thought of as subsumed under two respective event types, *A* and *B*, which are related by a generalization that *all* events of type *A* are followed by events of type *B*, can *a* be said to 'cause' *b*. When the events *a* and *b* are

[56] Writing this sentence, I find myself prey to an appropriate fear that (some) experts in Hume and Berkeley will not approve of some particular thing that I say about these philosophers here. I have made no careful study of them for the purpose of this paper. Rather a crude and fairly conventional account of the 'rough outlines' of their views is used for purposes of comparison with Wittgenstein.

considered by themselves alone, no causal notions are applicable. This Humean conclusion might be called: the impossibility of private causation.

Can one reasonably protest: surely there is nothing the event *a* can do with the *help* of other events of the same type that it cannot do by itself! Indeed, to say that *a*, by itself, is a sufficient cause of *b* is to say that, had the rest of the universe been removed, *a* still would have produced *b*! Intuitively this may well be so, but the intuitive objection ignores Hume's sceptical argument. The whole point of the sceptical argument is that the common notion of one event 'producing' another, on which the objection relies, is in jeopardy. It appears that there is no such relation as 'production' at all, that the causal relation is fictive. After the sceptical argument has been seen to be unanswerable on its own terms, a sceptical solution is offered, containing all we can salvage of the notion of causation. It just is a feature of this analysis that causation makes no sense when applied to two isolated events, with the rest of the universe removed. Only inasmuch as these events are thought of as instances of event types related by a regularity can they be thought of as causally connected. If two particular events were somehow so *sui generis* that it was logically excluded that they be placed under any (plausibly natural) event types, causal notions would not be applicable to them.

Of course I am suggesting that Wittgenstein's argument against private language has a structure similar to Hume's argument against private causation. Wittgenstein also states a sceptical paradox. Like Hume, he accepts his own sceptical argument and offers a 'sceptical solution' to overcome the appearance of paradox. His solution involves a sceptical interpretation of what is involved in such ordinary assertions as "Jones means addition by '+'." The impossibility of private language emerges as a corollary of his sceptical solution of his own paradox, as does the impossibility of 'private causation' in Hume. It turns out that the sceptical solution does not allow us to speak of a single individual,

considered by himself and in isolation, as ever meaning anything. Once again an objection based on an intuitive feeling that no one else can affect what I mean by a given symbol ignores the sceptical argument that undermines any such naive intuition about meaning.

I have said that Wittgenstein's solution to his problem is a sceptical one. He does not give a 'straight' solution, pointing out to the silly sceptic a hidden fact he overlooked, a condition in the world which constitutes my meaning addition by 'plus'. In fact, he agrees with his own hypothetical sceptic that there is no such fact, no such condition in either the 'internal' or the 'external' world. Admittedly, I am expressing Wittgenstein's view more straightforwardly than he would ordinarily allow himself to do. For in denying that there is any such fact, might we not be expressing a philosophical thesis that doubts or denies something everyone admits? We do not wish to doubt or deny that when people speak of themselves and others as meaning something by their words, as following rules, they do so with perfect right. We do not even wish to deny the propriety of an ordinary use of the phrase 'the fact that Jones meant addition by such-and-such a symbol', and indeed such expressions do have perfectly ordinary uses. We merely wish to deny the existence of the 'superlative fact' that philosophers misleadingly attach to such ordinary forms of words, not the propriety of the forms of words themselves.

It is for this reason that I conjectured above (p. 5), that Wittgenstein's professed inability to write a work with conventionally organized arguments and conclusions stems at least in part, not from personal and stylistic proclivities, but from the nature of his work. Had Wittgenstein – contrary to his notorious and cryptic maxim in §128 – stated the outcomes of his conclusions in the form of definite theses, it would have been very difficuilt to avoid formulating his doctrines in a form that consists in apparent sceptical denials of our ordinary assertions. Berkeley runs into similar difficulties. Partly he avoids them by stating his thesis as the denial of the existence of 'matter', and claiming that 'matter' is a bit of philosophical

jargon, not expressive of our common sense view. Nevertheless he is forced at one point to say – apparently contrary to his usual official doctrine – that he denies a doctrine 'strangely prevailing amongst men'.[57] If, on the other hand, we do not state our conclusions in the form of broad philosophical theses, it is easier to avoid the danger of a denial of any ordinary belief, even if our imaginary interlocutor (e.g. §189; see also §195)[58] accuses us of doing so. Whenever our opponent insists on the perfect propriety of an ordinary form of expression (e.g. that 'the steps are determined by the formula', 'the future application is already present'), we can insist that if these expressions are properly understood, we agree. The danger comes when we try to give a precise formulation of exactly what it is that we *are* denying – *what* 'erroneous interpretation' our opponent is placing on ordinary means of expression. It may be hard to do this without producing yet another statement that, we must admit, is *still* 'perfectly all right, properly understood'.[59]

So Wittgenstein, perhaps cagily, might well disapprove of the straightforward formulation given here. Nevertheless I choose to be so bold as to say: Wittgenstein holds, with the

[57] Berkeley, *The Principles of Human Knowledge*, §4. Of course Berkeley might mean that the prevalence of the doctrine stems from the influence of philosophical theory rather than common sense, as indeed he asserts in the next section.

[58] §189: "But *are* the steps then *not* determined by the algebraic formula?" In spite of Wittgenstein's interpretation within his own philosophy of the ordinary phrase "the steps are determined by the formula", the impression persists that the interlocutor's characterization of his view is really correct. See §195: "But I don't mean that what I do now (in grasping a sense) determines the future use *causally* and as a matter of experience, but that in a *queer* way, the use itself is in some sense present," which are the words of the interlocutor, and the bland reply, "But of course it is, 'in *some* sense'! Really the only thing wrong with what you say is the expression "in a queer way". The rest is all right; and the sentence only seems queer when one imagines a different language-game for it from the one in which we actually use it."

[59] An example of the kind of tension that can be involved appeared already above – see pp. 49–51 and note 33.

sceptic, that there is no fact as to whether I mean plus or quus. But if this is to be conceded to the sceptic, is this not the end of the matter? What *can* be said on behalf of our ordinary attributions of meaningful language to ourselves and to others? Has not the incredible and self-defeating conclusion, that all language is meaningless, already been drawn?

In reply we must say something about the change in Wittgenstein's philosophy of language from the *Tractatus* to the *Investigations*. Although in detail the *Tractatus* is among the most difficult of philosophical works, its rough outlines are well known. To each sentence there corresponds a (possible) fact. If such a fact, obtains, the sentence is true; if not, false. For atomic sentences, the relation between a sentence and the fact it alleges is one of a simple correspondence or isomorphism. The sentence contains names, corresponding to objects. An atomic sentence is itself a fact, putting the names in a certain relation; and it says that (there is a corresponding fact that) the corresponding objects are in the same relation. Other sentences are (finite or infinite) truth-functions of these. Even though the details of this theory have struck some as an implausible attempt to give natural language a chimerical *a priori* structure based on logical analysis alone, similar ideas, often advanced without any specific influence from the *Tractatus*, are much alive today.[60]

[60] Donald Davidson's influential and important theory of natural language has many features in common with the *Tractatus*, even if the underlying philosophy is different. Davidson argues that some simple, almost *a priori* considerations (not requiring detailed empirical investigation of specific natural languages) put strong constraints on the form of a theory of meaning for natural languages (it must be a finitely axiomatized Tarski-style theory of truth conditions). (Although the *form* of a theory is determined without detailed empirical investigation, for a particular language the specific theory adopted is supposed to require detailed empirical support.) The fact that a theory of meaning must have this form, it is argued, puts strong constraints on the logical form, or deep structure, of natural language – very probably that it ought to be close to classical extensional first order logic. All these ideas are close to the spirit of the *Tractatus*. In particular, like the *Tractatus*, Davidson holds (i) that truth conditions are a key element in a theory of language; (ii) that the

The simplest, most basic idea of the *Tractatus* can hardly be dismissed: a declarative sentence gets its meaning by virtue of its *truth conditions*, by virtue of its correspondence to facts that must obtain if it is true. For example, "the cat is on the mat" is understood by those speakers who realize that it is true if and only if a certain cat is on a certain mat; it is false otherwise. The presence of the cat on the mat is a fact or condition-in-the-world that would make the sentence true (express a truth) if it obtained.

So stated, the *Tractatus* picture of the meaning of declarative

uncovering of a hidden deep structure of language is crucial to a proper theory of interpretation; (iii) that the form of the deep structure is constrained in advance by theoretical, quasi-logical considerations; (iv) that, in particular, the constraints show that the deep structure has a logical form close to that of a formal language of symbolic logic; (v) that, in particular, sentences are built up from 'atoms' by logical operators; (vi) that, in particular, the deep structure of natural language is extensional in spite of the misleading appearances of surface structure. All these ideas of the *Tractatus* are repudiated in the *Investigations*, which is hostile to any attempt to analyze language by uncovering a hidden deep structure. In this last respect, modern transformational linguistics, since Noam Chomsky, has been closer to the *Tractatus* than to the *Investigations*. (But for transformational grammarians, even the form of the theory is established by specific empirical considerations requiring detailed investigation of specific natural languages.)

See also the programs of the linguists who called themselves 'generative semanticists' and of Richard Montague. Of course many of the ideas of the *Tractatus*, or of 'logical atomism', have not been revived in any of these theories.

(Note: In recent transformational linguistics, 'deep structure' has a specific technical meaning. 'Generative semanticists' made the repudiation of 'deep structure' a key plank of their platform. In the preceding, it is best to take 'deep structure' in the general sense of 'underlying' structure. Anyone whose theory of language leads him to applaud the doctrine of *Tractatus* 4.002 – that the understanding of language involves countless tacit conventions, invisible to the naked eye, that disguise form – believes in deep structure in this broad sense. 'Deep structure' in the specific sense was a special theory of deep structure thus broadly defined; that is one reason why it was an appropriate term. Most recent linguistic theories that rejected 'deep structure' in the specific sense accepted it in the broader sense.)

sentences may seem not only natural but even tautological. Nonetheless, as Dummett says, "the *Investigations* contains implicitly a rejection of the classical (realist) Frege–*Tractatus* view that the general form of explanation of meaning is a statement of the truth conditions".[61] In the place of this view, Wittgenstein proposes an alternative rough general picture. (To call it an alternative *theory* probably goes too far. Wittgenstein disclaims (§65) any intent of offering a general account of language to rival that of the *Tractatus*. Rather we have different activities related to each other in various ways.) Wittgenstein replaces the question, "What must be the case for this sentence to be true?" by two others: first, "Under what conditions may this form of words be appropriately asserted (or denied)?"; second, given an answer to the first question, "What is the role, and the utility, in our lives of our practice of asserting (or denying) the form of words under these conditions?"

Of course Wittgenstein does not confine himself to declarative sentences, and hence to assertion and denial, as I have just done. On the contrary, any reader of the earlier parts of *Philosophical Investigations* will be aware that he is strongly concerned to deny any special primacy to assertion, or to sentences in the indicative mood. (See his early examples "Slab!", "Pillar!", etc.) This in itself plays an important role in his repudiation of the classical realist picture. Since the indicative mood is not taken as in any sense primary or basic, it becomes more plausible that the linguistic role even of utterances in the indicative mood that superficially look like assertions need not be one of 'stating facts'.[62] Thus, if we speak properly, we should not speak of conditions of 'asser-

[61] Dummett, "Wittgenstein's Philosophy of Mathematics," p. 348 in original; reprinted in Pitcher (ed.), *Wittgenstein: The Philosophical Investigations*, pp. 446–7.

[62] See, for example, §304, where Wittgenstein is dealing with sensation language: "The paradox disappears only if we make a radical break with the idea that language . . . always serves the same purpose: to convey thoughts – which may be about houses, pains, good and evil, or anything else you please."

tion', but rather, more generally, of the conditions when a move (a form of linguistic expression) is to be made in the 'language game'. If, however, we allow ourselves to adopt an oversimplified terminology more appropriate to a special range of cases, we can say that Wittgenstein proposes a picture of language based, not on *truth conditions*, but on *assertability conditions* or *justification conditions*:[63] under what circumstances are we allowed to make a given assertion? Pictures, indeed explicit theories, of this kind are hardly unknown before

[63] Speaking of 'justification conditions' does not suggest the primacy of the indicative mood as much as 'assertability conditions', but it has its own drawbacks. For Wittgenstein, there is an important class of cases where a use of language properly has no independent justification other than the speaker's inclination to speak thus on that occasion (e.g. saying that one is in pain). In such cases, Wittgenstein says (§289), "To use a word without a justification (*Rechtfertigung*) does not mean to use it *zu Unrecht*." Anscombe's translation of '*zu Unrecht*' is not consistent. In her translation of *Philosophical Investigations*, §289, she translates it 'without right'. However, in her translation of *Remarks on the Foundations of Mathematics*, v, §33 [VII, §40], where almost exactly the same German sentence occurs, she translates it as 'wrongfully'. The German–English dictionary I have at hand (Wildhagen-Heraucourt, Brandstetter Verlag, Wiesbaden, and Allen and Unwin, London, 6th ed., 1962), translates '*zu Unrecht*' as 'unjustly, unfairly'; '*Unrecht*' in general is an 'injustice' or a 'wrong'. All this is reasonably consistent with 'wrongfully' but gives little support to 'without right', even though the idea that we have a 'right' to use a word in certain circumstances without 'justification' ('*Rechtfertigung*') is obviously in harmony with the point Wittgenstein is trying to make. However, by '*zu Unrecht*' Wittgenstein seems to mean that the use of a word without independent justification need not be a 'wrongful' use of the word – one without proper epistemic or linguistic support. On the contrary, it is essential to the workings of our language that, in some cases, such a use of language is perfectly proper. When we use the terminology of 'justification conditions', we must construe them to include such cases (where Wittgenstein would say there is no 'justification'). (Simply 'wrongly', might be a more idiomatic translation than 'wrongfully'. 'Without right' sounds to me too much as if a difficult new technical term is being introduced. The point is that '*zu Unrecht*', being a fairly ordinary German expression, should not be rendered so as to appear to be an unusual technical expression in English.) See also pp. 87–8 and note 75 below.

Wittgenstein and probably influenced him. The positivist verification theory of meaning is one of this kind. So, in a more special context, is the intuitionist account of mathematical statements. (The classical mathematician's emphasis on truth conditions is replaced by an emphasis on provability conditions.) But of course Wittgenstein's rough picture should not be identified with either of these. Its second component is distinct: granted that our language game permits a certain 'move' (assertion) under certain specifiable conditions, what is the role in our lives of such permission? Such a role must exist if this aspect of the language game is not to be idle.

Wittgenstein's alternative picture of language is already clearly suggested in the very first section of *Philosophical Investigations*. Many philosophers of mathematics – in agreement with the Augustinian conception of 'object and name' – ask such questions as, "What entities ('numbers') are denoted by numerals? What relations among these entities ('facts') correspond to numerical statements?" (Nominalistically inclined philosophers would counter, sceptically, "Can we really believe that there are such entities?") As against such a 'Platonist' conception of the problem, Wittgenstein asks that we discard any *a priori* conceptions and *look* ("Don't think, look!") at the circumstances under which numerical assertions are actually uttered, and at what roles such assertions play in our lives.[64] Suppose I go to the grocer with a slip marked 'five

[64] In some ways Frege can be taken to be the target here. It is he who insists on regarding numbers as *objects*, and on asking about the nature of these objects (even insisting that we can ask whether Julius Caesar is a number or not). On the other hand, the famous contextual principle of *Grundlagen der Arithmetik* (that one should ask for the signification of a sign only in the context of a sentence) and his emphasis in particular on asking how numerical expressions are actually applied is in the spirit of Wittgenstein's discussion. Perhaps the best conception of Wittgenstein's relation to Frege here is to say that Wittgenstein would regard the spirit of Frege's contextual principle as sound but would criticize Frege for using 'name of an object' as a catch-all for uses of language that are 'absolutely unlike' (§10).

red apples', and he hands over apples, reciting by heart the numerals up to five and handing over an apple as each numeral is intoned. It is under circumstances such as these that we are licensed to make utterances using numerals; the role and utility of such a license is obvious. In §§8–10, Wittgenstein imagines the letters of the alphabet, recited in alphabetical order, used in a miniature language game, just as the numbers are in this example. We have little inclination to wonder about the nature of the entities 'denoted' by the letters of the alphabet. Nevertheless, if they are used in the way described, they can properly be said to 'stand for numbers'. Indeed, to say words stand for (natural) numbers *is* to say that they are used as numerals, that is, used in the way described. Nevertheless the legitimacy, in its own way, of the expression 'stand for numbers' should not lead us to think of numerals as similar to expressions such as 'slab', 'pillar', and the like, except that the entities 'denoted' are not spatio-temporal. If the use of the expression 'stands for numbers' misleads in this way, it would be best to think in terms of another terminology, say, that an expression 'plays the role of a numeral'. This role, as Wittgenstein describes it, is plainly in strong *contrast* with the role of such expressions as 'slab', 'pillar', 'block', in the language games he describes in his early sections. (See §10.)

The case is a fine example of various aspects of Wittgenstein's technique in the *Investigations*. An important view in the philosophy of mathematics is suggested briefly almost *en passant*, almost hidden in a general discussion of the nature of language and 'language games'.[65] In the style discussed above,

[65] Paul Benacerraf, in "What Numbers Could Not Be," *The Philosophical Review*, vol. 74 (1963), pp. 47–73, see especially pp. 71–2, concludes with suggestions strikingly similar to Wittgenstein's though much of the preceding argumentation has no direct parallel in Wittgenstein. It is possible that one reason the resemblance of the views to those of a fairly well-known portion of the *Investigations* was not noticed is the *en passant* way Wittgenstein introduces the issue in the philosophy of mathematics in the context of a more general discussion. (Although I do not take it upon myself to criticize Wittgenstein in this essay, it seems to me that a great deal of further work must be done if one wishes to defend Wittgenstein's position here, since mathematics involves much more by way of apparently treating numbers as entities than can be covered by the simple case of counting. Perhaps some later authors can be interpreted as attempting to carry out such a project, but it is not my task to discuss these issues here.)

Wittgenstein suggests that such an expression as 'stands for a number' is in order, but is dangerous if it is taken to make a certain metaphysical suggestion. In the sense this is intended by 'Platonists', one suspects him of *denying* that numerals stand for entities called 'numbers'. Most important for the present purpose, the case exemplifies the central questions he wishes to ask about the use of language. Do not look for 'entities' and 'facts' corresponding to numerical assertions, but look at the circumstances under which utterances involving numerals are made, and the utility of making them under these circumstances.

Now the replacement of truth conditions by justification conditions has a dual role in the *Investigations*. First, it offers a new approach to the problems of how language has meaning, contrasted with that of the *Tractatus*. But second, it can be applied to give an account of assertions about meaning themselves, regarded as assertions *within* our language. Recall Wittgenstein's sceptical conclusion: no facts, no truth conditions, correspond to statements such as "Jones means addition by '+'." (The present remarks about meaning and use do not in themselves provide such truth conditions. According to them, Jones now means addition by '+' if he presently intends to use the '+' sign in one way, quaddition if he intends to use it another way. But nothing is said to illuminate the question as to the nature of such an intention.)

Now if we suppose that facts, or truth conditions, are of the essence of meaningful assertion, it will follow from the sceptical conclusion that assertions that anyone ever means anything are meaningless. On the other hand, if we apply to these assertions the tests suggested in *Philosophical Investigations*, no such conclusion follows. All that is needed to legitimize assertions that someone means something is that

there be roughly specifiable circumstances under which they are legitimately assertable, and that the game of asserting them under such conditions has a role in our lives. No supposition that 'facts correspond' to those assertions is needed.

I would therefore give the following rough structure to *Philosophical Investigations* (but the breaks between parts are not sharp and to an extent are arbitrary). §§1–137 give Wittgenstein's preliminary refutation of the *Tractatus* theory of language, and suggest the rough picture he intends to put in its place. These sections come first for more than one reason. First, Wittgenstein himself once found the *Tractatus* theory natural and inevitable – Malcolm says that even in his later period he regarded it as the *only* alternative to his later work[66] – and sometimes he writes as if the reader will naturally be inclined to the *Tractatus* theory unless he personally intervenes to prevent it. Thus the initial sections contain a refutation, not only of the most basic and apparently inevitable theories of the *Tractatus* (such as meaning as stating facts), but also of many of its more special doctrines (such as that of a special realm of 'simples').[67] Wittgenstein's contrast in these initial sections between his new way of looking at matters and his old way of thinking ranges from such special views of the *Tractatus* to the nature of philosophy. This first aspect of the initial sections has, I think, been clear to most readers. Less obvious is a second aspect. The sceptical paradox is the fundamental problem of *Philosophical Investigations*. If Wittgenstein is right, we cannot begin to solve it if we remain in the grip of the natural presupposition that meaningful declarative sentences

[66] See Norman Malcolm, *Ludwig Wittgenstein: A Memoir*, with a biographical sketch by G. H. von Wright (Oxford University Press, London, 1958), p. 69.

[67] Although Wittgenstein's concern in these initial sections is primarily with his own earlier way of thinking, of course he is concerned as well with related views (the 'object and name' model of language, the picture of sentences 'as corresponding to facts', etc.) in other writers, even though these writers may have views that differ in detail from those of the *Tractatus*. He wishes to relate the discussion to larger issues as well as to his own specific views.

must purport to correspond to facts; if this is our framework, we can only conclude that sentences attributing meaning and intention are themselves meaningless. Whether or not Wittgenstein is right in thinking that the entire *Tractatus* view is a consequence of natural and apparently inevitable presuppositions, he is surely right about this fundamental part of it. The picture of correspondence-to-facts must be cleared away before we can begin with the sceptical problem.

Sections 138–242 deal with the sceptical problem and its solution. These sections – the central sections of *Philosophical Investigations* – have been the primary concern of this essay. We have not yet looked at the solution of the problem, but the astute reader will already have guessed that Wittgenstein finds a useful role in our lives for a 'language game' that licenses, under certain conditions, assertions that someone 'means such-and-such' and that his present application of a word 'accords' with what he 'meant' in the past. It turns out that this role, and these conditions, involve reference to a community. They are inapplicable to a single person considered in isolation. Thus, as we have said, Wittgenstein rejects 'private language' as early as §202.

The sections following §243 – the sections usually called 'the private language argument' – deal with the *application* of the general conclusions about language drawn in §§138–242 to the problem of sensations. The sceptical conclusion about rules, and the attendant rejection of private rules, is hard enough to swallow in general, but it seems especially unnatural in two areas. The first is mathematics, the subject of most of the preceding discussion in the present essay (and of much of Wittgenstein's in §§138–242). Do I not, in elementary mathematics, grasp rules such as that for addition, which determine all future applications? Is it not in the very nature of such rules that, once I have grasped one, I have no future choice in its application? Is not any questioning of these assertions a questioning of mathematical proof itself? And is not the grasping of a mathematical rule the solitary achievement of each mathematician independent of any interaction

with a wider community? True, others may have taught me the concept of addition, but they acted only as heuristic aids to an achievement – the 'grasping of the concept' of addition – that puts me in a special relation to the addition function. Platonists have compared the grasping of a concept to a special sense, analogous to our ordinary sensory apparatus but percipient of higher entities. But the picture does not require a special Platonic theory of mathematical objects. It depends on the observation – apparently obvious on any view – that in grasping a mathematical rule I have achieved something that depends only on my own inner state, and that is immune to Cartesian doubt about the entire external material world.[68]

Now another case that seems to be an obvious counter-example to Wittgenstein's conclusion is that of a sensation, or mental image. Surely I can identify these after I have felt them, and any participation in a community is irrelevant! Because these two cases, mathematics and inner experience, seem so obviously to be counterexamples to Wittgenstein's view of rules, Wittgenstein treats each in detail. The latter case is treated in the sections following §243. The former case is treated in remarks that Wittgenstein never prepared for publication, but which are excerpted in *Remarks on the Foundations of Mathematics* and elsewhere. He thinks that only if we overcome our strong inclination to ignore his general conclusions about rules can we see these two areas rightly. For this reason, the conclusions about rules are of crucial importance both to the philosophy of mathematics and to the philosophy of mind. Although in his study of sensations in

[68] Although Wittgenstein's views on mathematics were undoubtedly influenced by Brouwer, it is worth noting here that Brouwer's intuitionist philosophy of mathematics is, if anything, even more solipsistic than its traditional 'Platonist' rival. According to this conception, mathematics can be idealized as the isolated activity of a single mathematician ('creating subject') whose theorems are assertions about his own mental states. The fact that mathematicians form a community is irrelevant for theoretical purposes. (Indeed, Brouwer himself is said to have held mysterious 'solipsistic' views that communication is impossible. The point would remain even if we left these aside.)

§243 onward he does not simply *cite* his general conclusions but argues this special case afresh (he does the same for mathematics elsewhere), we will only increase our difficulties in understanding an already difficult argument if we call §243 onward 'the private language argument' and study it in isolation from the preceding material. Wittgenstein had a definite plan of organization when he placed this discussion where it is.

Of course the division is not sharp. The initial 'anti-*Tractatus*' sections contain several anticipations of the 'paradox' of §§138–242,[69] and even of its solution. Sections 28–36 and sections 84–9 are examples. Even the very first section of the *Investigations* can be read, with hindsight, as anticipating the problem.[70] Nevertheless these anticipations, being cryptic allusions to the problem in the context of the problems of earlier discussion, do not fully develop the paradox and often elide the main point into other subsidiary ones.

Consider first the anticipation in sections 84–9, especially section 86, where Wittgenstein introduces the ambiguity of rules and the possibility of an infinite regress of 'rules to interpret rules'. Knowing the central problem of *Philosophical Investigations*, it is easy to see that in these sections Wittgenstein is concerned to bring out this problem, and even to allude to part of his approach to a solution (end of §87: "The sign post is in order if, in normal circumstances, it serves its purpose"). In the context, however, Wittgenstein shades his deep paradox into a much more straightforward point – that typically

[69] Barry Stroud emphasized this fact to me, though the responsibility for the examples and exposition in the following paragraphs is my own.

[70] See: "But how does he know where and how he is to look up the word 'red' and what he is to do with the word 'five'? – Well, I assume that he *acts* as I described. Explanations come to an end somewhere." (§1) In hindsight, this is a statement of the basic point that I follow rules 'blindly', without any justification for the choice I make. The suggestion in the section that nothing is wrong with this situation, provided that my use of 'five', 'red', etc. fits into a proper system of activities in the community, anticipates Wittgenstein's sceptical solution, as expounded below.

uses of language do not give a precise determination of their application in all cases. (See the discussion of names in §79 – "I use the name . . . without a *fixed* meaning"; of the 'chair' (?) in §80; 'Stand roughly here' in §88.) It is true, as Wittgenstein says, that his paradox shows, among other things, that every explanation of a rule could conceivably be misunderstood, and that in this respect the most apparently precise use of language does not differ from 'rough' or 'inexact', or 'open-textured' uses. Nevertheless, surely the real point of Wittgenstein's paradox is not that the rule of addition is somehow *vague*, or leaves some cases of its application undetermined. On the contrary, the word 'plus' denotes a function whose determination is *completely* precise – in this respect it does *not* resemble the vague notions expressed by 'large', 'green', and the like. The point is the sceptical problem, outlined above, that anything in my head leaves it undetermined *what* function 'plus' (as I use it) denotes (plus or quus), what 'green' denotes (green or grue), and so on. The ordinary observation, made in abstraction from any scepticism about the meaning of 'green', that the property of greenness is itself only vaguely defined for some cases, is at best distantly related. In my opinion, Wittgenstein's sceptical arguments in no way show, in this sense, that the addition function is only vaguely defined. The addition function – as Frege would emphasize – yields one precise value for each pair of numerical arguments. This much is a theorem of arithmetic. The sceptical problem indicates no vagueness in the *concept* of addition (in the way there *is* vagueness in the concept of greenness), or in the word 'plus', *granting* it its usual meaning (in the way the word 'green' *is* vague). The sceptical point is something else.[71]

[71] Though perhaps vagueness, in the ordinary sense, enters into Wittgenstein's puzzle in this way: when a teacher *introduces* such a word as 'plus' to the learner, if he does not reduce it to more 'basic', previously learned concepts, he introduces it by a finite number of examples, plus the instructions: "Go on in the same way!" The last clause may indeed be regarded as vague, in the ordinary sense, though our grasp of the most precise concept depends on it. This type of vagueness *is* intimately connected with Wittgenstein's paradox.

In the sections under discussion, Wittgenstein is arguing that *any* explanation *may* fail of its purpose: if it does not in fact fail, it may work perfectly, even if the concepts involved violate the Fregean requirement of 'sharp boundaries' (§71). See §88: "If I tell someone "Stand roughly here" may not this explanation work perfectly? And cannot every other one fail too?" At least two issues are involved here: the propriety of vagueness, of violations of the Fregean requirement (actually Wittgenstein questions whether this requirement, in an absolute sense, is well-defined); and an adumbration of the sceptical paradox of the second portion (§§138–242) of the *Investigations*. In its present context, the paradox, briefly foreshadowed, is not clearly distinguished from the other considerations about vagueness and sharp boundaries. The real development of the problem is yet to come.

Similar remarks apply to the discussion of ostensive definition in §§28–36, which is part of a larger discussion of naming, one of the important topics for the first portion (§§1–137) of the *Investigations*. Wittgenstein emphasizes that ostensive definitions are always in principle capable of being misunderstood, even the ostensive definition of a color word such as 'sepia'. How someone understands the word is exhibited in the way someone goes on, "the use that he makes of the word defined". One may go on in the right way given a purely minimal explanation, while on the other hand one may go on in another way no matter how many clarifications are added, since these too can be misunderstood (a rule for interpreting a rule again; see especially §§28–9).

Much of Wittgenstein's argument is directed against the view of a special, qualitatively unique experience of understanding the ostensive definition in the right way (§§33–6). Once again Wittgenstein's real point, here in the context of naming and ostensive definition, is the sceptical paradox. The case of ostensive definition of a color ('sepia') has a special connection with the so-called 'private language argument', as developed for sensations in §§243ff. Here too, however, the argument is adumbrated so briefly, and is so much embedded

in a context of other issues, that at this stage of the argument the point can easily be lost.[72]

Yet another feature of the situation indicates how the ideas can be connected in a way that cuts across the indicated divisions of *Philosophical Investigations*. The first part (up to §137), as we have said, criticizes Wittgenstein's earlier picture of the nature of language and attempts to suggest another. Since Wittgenstein's sceptical solution of his paradox is possible only given his later conception of language and is ruled out by the earlier one, the discussion in the second part (§§138–242) is dependent on that of the first. The point to be made here is that, at the same time, the second part is important for an ultimate understanding of the first. Wittgen-

[72] In these sections, Wittgenstein does not cite examples like 'grue' or 'quus' but begins by emphasizing the ordinary possibilities for misunderstanding an ostensive definition. Many philosophers who have been influenced by Wittgenstein have happened also to be attracted to the idea that an act of ostension is ill defined unless it is accompanied by a sortal ('the entity I am pointing to' versus 'the color I am pointing to', 'the shape . . .', 'the table . . .', etc.). Then morals regarding naming and identity (as associated with 'sortal terms') are drawn from this fact. I have the impression that many of these philosophers would interpret Wittgenstein's §§28–9 as making the same point. (See, e.g., M. Dummett, *Frege* (Duckworth, London, 1973, xxv + 698 pp.), pp. 179–80, and frequently elsewhere.) However, it seems clear to me that the *main* point of these sections is almost the exact opposite. It should be clear from reading §29 that the idea of adding a sortal ("This *number* is called 'two'") is introduced by Wittgenstein's imaginary interlocutor. As against this, Wittgenstein replies that the point is in a sense correct, but that the original ostensive definition – without a sortal – is perfectly legitimate provided that it leads the learner to apply such a word as 'two' correctly in the future, while even if the sortal term is added, the possibility of future misapplication is not removed, since the sortal too may be interpreted incorrectly (and this problem cannot be removed by further explanations). Really there are two separable issues, as in the case of §§84–9. One issue is analogous to the one about vagueness in §§84–9: that an ostensive definition without an accompanying sortal is vague. The other, which clearly is the main point, is Wittgenstein's sceptical problem, presented here in terms of the possibility of misunderstanding an ostensive definition.

stein's earlier work had taken for granted a natural relation of interpretation between a thought in someone's mind and the 'fact' it 'depicts'. The relation was supposed to consist in an isomorphism between one fact (the fact that mental elements are arranged in a certain way) and another (the fact-in-the-world 'depicted'). Some of Wittgenstein's attack on this earlier idea is developed in the first part through a criticism of the notion, crucial to the *Tractatus* theory of isomorphism, of a unique decomposition of a complex into its 'ultimate' elements (see, for example, §§47–8). Clearly, however, the paradox of the second part of the *Investigations* constitutes a powerful critique of any idea that 'mental representations' uniquely correspond to 'facts', since it alleges that the components of such 'mental representations' do not have interpretations that can be 'read off' from them in a unique manner. So *a fortiori* there is no such unique interpretation of the mental 'sentences' containing them as 'depicting' one 'fact' or another.[73] In this way the relationship between the first and the second portions of the *Investigations* is reciprocal. In order for Wittgenstein's sceptical solution of his paradox to be intelligible, the 'realistic' or 'representational' picture of language must be undermined by another picture (in the first part). On the other hand, the paradox developed in the second part, antecedently to its solution, drives an important final nail (perhaps the crucial one) into the coffin of the representational picture.[74] No doubt this is one reason Wittgenstein introduces foreshadowings of the paradox already in the sections of the first part. But it also illustrates that the structural divisions I have indicated in *Philosophical Investigations* are not sharp. The investigation goes 'criss cross in every direction' (preface).

[73] The criticisms of the earlier ideas about 'isomorphism' are thus criticisms of a special alleged way of obtaining a unique interpretation of a mental representation. For Wittgenstein, given his earlier views, criticisms of the notion of isomorphism are thus of obvious special importance as a stage setting for his paradox. They are relatively less important as such a stage setting for someone who is not working his way out of this special *milieu*.

[74] Michael Dummett emphasized this point to me, though the responsibility for the present formulation is my own.

Wittgenstein's sceptical solution concedes to the sceptic that no 'truth conditions' or 'corresponding facts' in the world exist that make a statement like "Jones, like many of us, means addition by '+'" true. Rather we should look at how such assertions are *used*. Can this be adequate? Do we not call assertions like the one just quoted 'true' or 'false'? Can we not with propriety precede such assertions with 'It is a fact that' or 'It is not a fact that'? Wittgenstein's way with such objections is short. Like many others, Wittgenstein accepts the 'redundancy' theory of truth: to affirm that a statement is true (or presumably, to precede it with 'It is a fact that . . .') is simply to affirm the statement itself, and to say it is not true is to deny it: ('*p*' is true = *p*). However, one might object: (a) that only utterances of certain forms are called 'true' or 'false' – questions, for example, are not – and these are so called precisely because they purport to state facts; (b) that precisely the sentences that 'state facts' can occur as components of truth-functional compounds and their meaning in such compounds is hard to explain in terms of assertability conditions alone. Wittgenstein's way with this is also short. We *call* something a proposition, and hence true or false, when in our language we apply the calculus of truth functions to it. That is, it is just a primitive part of our language game, not susceptible of deeper explanation, that truth functions are applied to certain sentences. For the present expository purpose it is worth noting that the sections in which he discusses the concept of truth (§§134–7) *conclude* the preliminary sections on the *Tractatus* and immediately *precede* the discussion of the sceptical paradox. They lay the final groundwork needed for that discussion.

Finally, we can turn to Wittgenstein's sceptical solution and to the consequent argument against 'private' rules. We have to see under what circumstances attributions of meaning are made and what role these attributions play in our lives. Following Wittgenstein's exhortation not to think but to look, we will not reason *a priori* about the role such statements *ought* to play; rather we will find out what circumstances *actually* license such assertions and what role this license *actually* plays. It is important to realize that we are *not* looking for necessary and sufficient conditions (truth conditions) for following a rule, or an analysis of what such rule-following 'consists in'. Indeed such conditions would constitute a 'straight' solution to the sceptical problem, and have been rejected.

First, consider what is true of one person considered in isolation. The most obvious fact is one that might have escaped us after long contemplation of the sceptical paradox. It holds no terrors in our daily lives; no one actually hesitates when asked to produce an answer to an addition problem! Almost all of us unhesitatingly produce the answer '125' when asked for the sum of 68 and 57, without any thought to the theoretical possibility that a quus-like rule might have been appropriate! And we do so without justification. Of course, if asked why we said '125', most of us will say that we added 8 and 7 to get 15, that we put down 5 and carried 1 and so on. But then, what will we say if asked why we 'carried' as we do? Might our past intention not have been that 'carry' meant *quarry*; where to 'quarry' is . . .? The entire point of the sceptical argument is that ultimately we reach a level where we act without any reason in terms of which we can justify our action. We act unhesitatingly but *blindly*.

This then is an important case of what Wittgenstein calls speaking without 'justification' ('*Rechtfertigung*'), but not 'wrongfully' ('*zu Unrecht*').[75] It is part of our language game of speaking of rules that a speaker may, without ultimately giving any justification, follow his own confident inclination that this way (say, responding '125') is the *right* way to

[75] See note 63. Note that in *Remarks on the Foundations of Mathematics*, V, §33 [VII, §40], Wittgenstein develops this point with respect to his general problem about rules, agreement, and identity, while the parallel passage in *Philosophical Investigations*, §289, is concerned with avowals of pain. This illustrates again the connection of Wittgenstein's ideas on sensation language with the general point about rules. Note also that the *RFM* passage is embedded in a context of the philosophy of mathematics. The connection of Wittgenstein's discussions of mathematics with his discussions of sensations is another theme of the present essay.

respond, rather than another way (e.g. responding '5'). That is, the 'assertability conditions' that license an individual to say that, on a given occasion, he ought to follow his rule this way rather than that, are, ultimately, that he does what he is inclined to do.

The important thing about this case is that, if we confine ourselves to looking at one person alone, his psychological states and his external behavior, this is as far as we can go. We can say that he acts confidently at each application of a rule; that he says – without further justification – that the way he acts, rather than some quus-like alternative, is *the* way to respond. There are no circumstances under which we can say that, even if he inclines to say '125', he *should* have said '5', or *vice versa*. By definition, *he* is licensed to give, without further justification, the answer that strikes him as natural and inevitable. Under what circumstances can he be wrong, say, following the wrong rule? No one else by looking at his mind and behavior alone can say something like, "He is wrong if he does not accord with his own past intentions"; the whole point of the sceptical argument was that there can be no facts about him in virtue of which he accords with his intentions or not. All we can say, if we consider a single person in isolation, is that our ordinary practice licenses him to apply the rule in the way it strikes him.

But of course this is *not* our usual concept of following a rule. It is by no means the case that, just because someone thinks he is following a rule, there is no room for a judgement that he is not really doing so. Someone – a child, an individual muddled by a drug – may think he is following a rule even though he is actually acting at random, in accordance with no rule at all. Alternatively, he may, under the influence of a drug, suddenly act in accordance with a quus-like rule changing from his first intentions. If there could be no justification for anyone to say of a person of the first type that his confidence that he is following some rule is misplaced, or of a person of the second type that he is no longer in accord with the rule that he previously followed, there would be little

content to our idea that a rule, or past intention, *binds* future choices. We are inclined to accept conditionals of such a rough type as, "If someone means addition by '+' then, if he remembers his past intention and wishes to conform to it, when he is queried about '68+57', he will answer '125'." The question is what substantive content such conditionals can have.

If our considerations so far are correct, the answer is that, if one person is considered in isolation, the notion of a rule as guiding the person who adopts it can have *no* substantive content. There are, we have seen, no truth conditions or facts in virtue of which it can be the case that he accords with his past intentions or not. As long as we regard him as following a rule 'privately', so that we pay attention to *his* justification conditions alone, all we can say is that he is licensed to follow the rule as it strikes him. This is why Wittgenstein says, "To think one is obeying a rule is not to obey a rule. Hence it is not possible to obey a rule 'privately'; otherwise thinking one was obeying a rule would be the same thing as obeying it." (§202)

The situation is very different if we widen our gaze from consideration of the rule follower alone and allow ourselves to consider him as interacting with a wider community. Others will then have justification conditions for attributing correct or incorrect rule following to the subject, and these will *not* be simply that the subject's own authority is unconditionally to be accepted. Consider the example of a small child learning addition. It is obvious that his teacher will not accept just any response from the child. On the contrary, the child must fulfill various conditions if the teacher is to ascribe to him mastery of the concept of addition. First, for small enough examples, the child must produce, almost all the time, the 'right' answer. If a child insists on the answer '7' to the query '2+3', and a '3' to '2+2', and makes various other elementary mistakes, the teacher will say to him, "You are not adding. Either you are computing another function" – I suppose he would not really talk quite this way to a child! – "or, more probably, you are as yet following no rule at all, but only giving whatever random

answer enters your head." Suppose, however, the child gets almost all 'small' addition problems right. For larger computations, the child can make more mistakes than for 'small' problems, but it must get a certain number right and, when it is wrong, it must recognizably be 'trying to follow' the proper procedure, not a quus-like procedure, even though it makes mistakes. (Remember, the teacher is not judging how accurate or *adept* the child is as an adder, but whether he can be said to be following the rule for adding.) Now, what do I mean when I say that the teacher judges that, for certain cases, the pupil must give the 'right' answer? I mean that the teacher judges that the child has given the same answer that he himself would give. Similarly, when I said that the teacher, in order to judge that the child is adding, must judge that, for a problem with larger numbers, he is applying the 'right' procedure even if he comes out with a mistaken result, I mean that he judges that the child is applying the procedure he himself is inclined to apply.

Something similar is true for adults. If someone whom I judge to have been computing a normal addition function (that is, someone whom I judge to give, when he adds, the same answer I would give), suddenly gives answers according to procedures that differ bizarrely from my own, then I will judge that something must have happened to him, and that he is no longer following the rule he previously followed. If this happens to him generally, and his responses seem to me to display little discernible pattern, I will judge him probably to have gone insane.

From this we can discern rough assertability conditions for such a sentence as "Jones means addition by 'plus'." *Jones* is entitled, subject to correction by others, provisionally to say, "I mean addition by 'plus'," whenever he has the feeling of confidence – "now I can go on!" – that he can give 'correct' responses in new cases; and *he* is entitled, again provisionally and subject to correction by others, to judge a new response to be 'correct' simply because it is the response he is inclined to give. These inclinations (both Jones's general inclination that

he has 'got it' and his particular inclination to give particular answers in particular addition problems) are to be regarded as primitive. They are not to be justified in terms of Jones's ability to interpret his own intentions or anything else. But Smith need *not* accept Jones's authority on these matters: *Smith* will judge Jones to mean addition by 'plus' only if he judges that Jones's answers to particular addition problems agree with those *he* is inclined to give, or, if they occasionally disagree, he can interpret Jones as at least following the proper procedure. (If Jones gives answers for very small problems disagreeing with those Smith is inclined to give, it will be difficult or impossible for Smith to interpret Jones as following the proper procedure. The same will hold if Jones's responses to larger problems are too bizarre to be errors in addition in the normal sense: for example, if he answers '5' to '68+57'.) If Jones consistently fails to give responses in agreement (in this broad sense) with Smith's, Smith will judge that he does not mean addition by 'plus'. Even if Jones did mean it in the past, the present deviation will justify Smith in judging that he has lapsed.

Sometimes Smith, by substituting some alternative interpretation for Jones's word 'plus', will be able to bring Jones's responses in line with his own. More often, he will be unable to do so and will be inclined to judge that Jones is not really following any rule at all. In all this, Smith's inclinations are regarded as just as primitive as Jones's. In no way does Smith test directly whether Jones may have in his head some rule agreeing with the one in Smith's head. Rather the point is that if, in enough concrete cases, Jones's inclinations agree with Smith's, Smith will judge that Jones is indeed following the rule for addition.

Of course if we were reduced to a babble of disagreement, with Smith and Jones asserting of each other that they are following the rule wrongly, while others disagreed with both and with each other, there would be little point to the practice just described. In fact, our actual community is (roughly) uniform in its practices with respect to addition. Any indi-

vidual who claims to have mastered the concept of addition will be judged by the community to have done so if his particular responses agree with those of the community in enough cases, especially the simple ones (and if his 'wrong' answers are not often *bizarrely* wrong, as in '5' for '68+57', but seem to agree with ours in *procedure*, even when he makes a 'computational mistake'). An individual who passes such tests is admitted into the community as an adder; an individual who passes such tests in enough other cases is admitted as a normal speaker of the language and member of the community. Those who deviate are corrected and told (usually as children) that they have not grasped the concept of addition. One who is an incorrigible deviant in enough respects simply cannot participate in the life of the community, and in communication.

Now Wittgenstein's general picture of language, as sketched above, requires for an account of a type of utterance not merely that we say under what conditions an utterance of that type can be made, but also what role and utility in our lives can be ascribed to the practice of making this type of utterance under such conditions. We say of someone else that he follows a certain rule when his responses agree with our own and deny it when they do not; but what is the utility of this practice? The utility is evident and can be brought out by considering again a man who buys something at the grocer's. The customer, when he deals with the grocer and asks for five apples, expects the grocer to count as he does, not according to some bizarre non-standard rule and so, if his dealings with the grocer involve a computation, such as '68+57', he expects the grocer's responses to agree with his own. Indeed, he may entrust the computation to the grocer. Of course the grocer may make mistakes in addition: he may even make dishonest computations. But as long as the customer attributes to him a grasp of the concept of addition, he expects that at least the grocer will not behave bizarrely, as he would if he were to follow a quus-like rule; and one can even expect that, in many cases, he will come up with the same answer the customer

would have given himself. When we pronounce that a child has mastered the rule of addition, we mean that we can entrust him to react as we do in interactions such as that just mentioned between the grocer and the customer. Our entire lives depend on countless such interactions, and on the 'game' of attributing to others the mastery of certain concepts or rules, thereby showing that we expect them to behave as we do.

This expectation is *not* infallibly fulfilled. It places a substantive restriction on the behavior of each individual, and is *not* compatible with just any behavior he may choose. (Contrast this with the case where we considered one person alone.) A deviant individual whose responses do not accord with those of the community in enough cases will not be judged, by the community, to be following its rules; he may even be judged to be a madman, following no coherent rule at all. When the community denies of someone that he is following certain rules, it excludes him from various transactions such as the one between the grocer and the customer. It indicates that it cannot rely on his behavior in such transactions.

We can restate this in terms of a device that has been common in philosophy, *inversion* of a conditional.[76] For example, it is important to our concept of causation that we accept some such conditional as: "If events of type *A* cause

[76] As will be seen immediately, inversion in this sense is a device for reversing priorities. William James summarized his famous theory of the emotions (*The Principles of Psychology*, Henry Holt & Co., New York, 1913, in 2 volumes; chapter 25 (vol. 2, 442–85), "The Emotions") by the assertion, ". . . the . . . rational statement is that we feel sorry because we cry . . . not that we cry . . . because we are sorry . . ." (p. 450). Many philosophies can be summed up crudely (no doubt, not really accurately) by slogans in similar form.: "We do not condemn certain acts because they are immoral; they are immoral because we condemn them." "We do not accept the law of contradiction because it is a necessary truth; it is a necessary truth because we accept it (by convention)." "Fire and heat are not constantly conjoined because fire causes heat; fire causes heat because they are constantly conjoined" (Hume). "We do not all say

events of type B, and if an event e of type A occurs, then an event e' of type B must follow." So put, it appears that acceptance of the conditional commits us to a belief in a nexus so that, given that the causal connection between event types obtains, the occurrence of the first event e necessitates (by fulfilling the antecedent of the conditional), that an event e' of type B must obtain. Humeans, of course, deny the existence of such a nexus; how do they read the conditional? Essentially they concentrate on the assertability conditions of a contrapositive form of the conditional. It is not that any antecedent conditions necessitate that some event e' must take place; rather the conditional commits us, whenever we know that an event e of type A occurs and is not followed by an event of type B, to deny that there is a causal connection between the two event types. If we did make such a claim, we must now withdraw it. Although a conditional is equivalent to its contrapositive, concentration on the contrapositive reverses our priorities. Instead of seeing causal connections as primary, from which observed regularities 'flow', the Humean instead sees the regularity as primary, and – looking at the matter contrapositively – observes that we withdraw a causal hypothesis when the corresponding regularity has a definite counter-instance.

A similar inversion is used in the present instance. It is essential to our concept of a rule that we maintain some such conditional as 'If Jones means addition by '+', then if he is asked for '68+57', he will reply '125'." (Actually many clauses should be added to the antecedent to make it strictly correct, but for present purposes let us leave it in this rough form.) As in the causal case, the conditional as stated makes it appear that

12+7=19 and the like because we all grasp the concept of addition; we say we all grasp the concept of addition because we all say 12+7=19 and the like" (Wittgenstein).

The device of inversion of a conditional in the text achieves the effect of reversing priorities in a way congenial to such slogans. Speaking for myself, I am suspicious of philosophical positions of the types illustrated by the slogans, whether or not they are so crudely put.

some mental state obtains in Jones that guarantees his performance of particular additions such as '68+57' – just what the sceptical argument denies. Wittgenstein's picture of the true situation concentrates on the contrapositive, and on justification conditions. If Jones does *not* come out with '125' when asked about '68+57', we cannot assert that he means addition by '+'. Actually, of course, this is not strictly true, because our formulation of the conditional is overly loose; other conditions must be added to the antecedent to make it true. As the conditional is stated, not even the possibility of computational error is taken into account, and there are many complications not easily spelled out. The fact remains that if we ascribe to Jones the conventional concept of addition, we do not expect him to exhibit a pattern of bizarre, quus-like behavior. By such a conditional we do not mean, on the Wittgensteinian view, that any state of Jones guarantees his correct behavior. Rather by asserting such a conditional we commit ourselves, if in the future Jones behaves bizarrely enough (and on enough occasions), no longer to persist in our assertion that he is following the conventional rule of addition.

The rough conditional thus expresses a restriction on the community's game of attributing to one of its members the grasping of a certain concept: if the individual in question no longer conforms to what the community would do in these circumstances, the community can no longer attribute the concept to him. Even though, when we play this game and attribute concepts to individuals, we depict no special 'state' of their minds, we do something of importance. We take them provisionally into the community, as long as further deviant behavior does not exclude them. In practice, such deviant behavior rarely occurs.

It is, then, in such a description of the game of concept attribution that Wittgenstein's sceptical solution consists. It provides both conditions under which we are justified in attributing concepts to others and an account of the utility of this game in our lives. In terms of this account we can discuss briefly three of Wittgenstein's key concepts.

First, *agreement*. The entire 'game' we have described – that the community attributes a concept to an individual so long as he exhibits sufficient conformity, under test circumstances, to the behavior of the community – would lose its point outside a community that generally agrees in its practices. If one person, when asked to compute '68+57' answered '125', another '5', and another '13', if there was no general agreement in the community responses, the game of attributing concepts to individuals – as we have described it – could not exist. In fact of course there is considerable agreement, and deviant quus-like behavior occurs rarely. Mistakes and disagreements do occur, but these are another matter. The fact is that, extreme cases of uneducability or insanity aside, almost all of us, after sufficient training, respond with roughly the same procedures to concrete addition problems. We respond unhesitatingly to such problems as '68 + 57', regarding our procedure as the only comprehensible one (see, e.g., §§219, 231, 238), and we *agree* in the unhesitating responses we make. On Wittgenstein's conception, such agreement is essential for our game of ascribing rules and concepts to each other (see §240).

The set of responses in which we agree, and the way they interweave with our activities, is our *form of life*. Beings who agreed in consistently giving bizarre quus-like responses would share in another form of life. By definition, such another form of life would be bizarre and incomprehensible to us. ("If a lion could talk, we could not understand him" (p. 223).) However, if we can imagine the abstract possibility of another form of life (and no *a priori* argument would seem to exclude it), the members of a community sharing such a quus-like form of life could play the game of attributing rules and concepts to each other as we do. Someone would be said, in such a community, to follow a rule, as long as he agrees in his responses with the (*quus*-like) responses produced by the members of *that* community. Wittgenstein stresses the importance of agreement, and of a shared form of life, for his solution to his sceptical problem in the concluding paragraphs of the central section of *Philosophical Investigations* (§§240–2; see also the discussion of agreement on pp. 225–7).

On Wittgenstein's conception, a certain type of traditional – and overwhelmingly natural – explanation of our shared form of life is excluded. We cannot say that we all respond as we do to '68+57' *because* we all grasp the concept of addition in the same way, that we share common responses to particular addition problems *because* we share a common concept of addition. (Frege, for example, would have endorsed such an explanation, but one hardly needs to be a philosopher to find it obvious and natural.) For Wittgenstein, an 'explanation' of this kind ignores his treatment of the sceptical paradox and its solution. There is no objective fact – that we all mean addition by '+', or even that a given individual does – that explains our agreement in particular cases. Rather our license to say of each other that we mean addition by '+' is part of a 'language game' that sustains itself only because of the brute fact that we generally agree. (Nothing about 'grasping concepts' guarantees that it will not break down tomorrow.) The rough uniformities in our arithmetical behavior may or may not some day be given an explanation on the neurophysiological level, but such an explanation is not here in question.[77] Note again the analogy with the Humean case. Naively, we may wish to explain the observed concomitance of fire and heat by a causal, heat-producing, 'power' in the fire. The Humean alleges that any such use of causal powers to explain the regularity is meaningless. Rather we play a language game that allows us to attribute such a causal power to the fire as

[77] Modern transformational linguistics, inasmuch as it explains all my specific utterances by my 'grasp' of syntactic and semantic rules generating infinitely many sentences with their interpretation, seems to give an explanation of the type Wittgenstein would not permit. For the explanation is *not* in terms of my actual 'performance' as a finite (and fallible) device. It is not a purely causal (neurophysiological) explanation in the sense explained in the text, see note 22 above. On the other hand, some aspects of Chomsky's views are very congenial to Wittgenstein's conception. In particular, according to Chomsky, highly species-specific constraints – a 'form of life' – lead a child to project, on the basis of exposure to a limited corpus of sentences, a variety of new sentences for new situations. There is no *a priori* inevitability in the child's going on in the way he does, other than that this is what the species does. As was already said in note 22, the matter deserves a more extended discussion.

long as the regularity holds up. The regularity must be taken as a brute fact. So too for Wittgenstein (p. 226): "What has to be accepted, the given, is . . . *forms of life*."[78]

Finally, *criteria*. The exact interpretation and exegesis of Wittgenstein's concept of a criterion has been the subject of much discussion among students of Wittgenstein's later work. Criteria play a fundamental role in Wittgenstein's philosophy of mind: "An 'inner process' stands in need of outward criteria" (§580). Often the necessity for criteria for mental concepts has been taken, both by advocates and critics of Wittgenstein's philosophy of mind, as a fundamental *premise* of

[78] Can we imagine forms of life other than our own, that is, can we imagine creatures who follow rules in bizarre quus-like ways? It seems to me that there may be a certain tension in Wittgenstein's philosophy here. On the one hand, it would seem that Wittgenstein's paradox argues that there is no *a priori* reason why a creature could not follow a quus-like rule, and thus in this sense we ought to regard such creatures as conceivable. On the other hand, it is supposed to be part of our very form of life that we find it natural and, indeed, inevitable that we follow the rule for addition in the particular way that we do. (See §231: "'But surely you can see . . .?' That is just the characteristic expression of someone who is under the compulsion of a rule.") But then it seems that we should be unable to understand 'from the inside' (cf. the notion of '*Verstehen*' in various German writers) how any creature could follow a quus-like rule. We could describe such behavior extensionally and behavioristically, but we would be unable to find it intelligible how the creature finds it natural to behave in this way. This consequence does, indeed, seem to go with Wittgenstein's conception of the matter.

Of course we can define the quus function, introduce a symbol for it, and follow the appropriate rule for computing its values. I have done so in this very essay. What it seems may be unintelligible to us is how an intelligent creature could get the very training we have for the addition function, and yet grasp the appropriate function in a quus-like way. If such a possibility were really completely intelligible to us, would we find it so inevitable to apply the plus function as we do? Yet this inevitability is an essential part of Wittgenstein's own solution to his problem.

The point is even stronger with respect to a term like 'green'. Can we grasp how someone could be presented with a number of green objects, and be told to apply the term 'green' just to 'things like these', and yet apply the term learnt as if it meant 'grue'? It would seem that if we find our own continuation to be inevitable, in some sense we cannot.

his private language argument. Critics have sometimes argued that it constitutes an undefended and indefensible verificationist assumption. Some advocates respond that if it is a verificationist premise of some sort, that form of verificationism is clearly correct.

It is not my present purpose to enter into the finer exegetical points involved in Wittgenstein's notion of a criterion,[79] but rather to sketch the role of the notion in the picture we have been developing. Wittgenstein's sceptical solution to his problem depends on agreement, and on checkability – on one person's ability to test whether another uses a term as he does. In our own form of life, how does this agreement come about? In the case of a term like 'table', the situation, at least in elementary cases, is simple. A child who says "table" or "That's a table" when adults see a table in the area (and does not do so otherwise) is said to have mastered the term 'table': he says "That's a table", based on his observation, in agreement with the usage of adults, based on their observation. That is, they say, "That's a table" under like circumstances, and confirm the correctness of the child's utterances.

How does agreement emerge in the case of a term for a sensation, say 'pain'? It is not as simple as the case of 'table'. When will adults attribute to a child mastery of the avowal "I am in pain"?[80] The child, if he learns the avowal correctly, will utter it when he feels pain and not otherwise. By analogy with the case of 'table', it would appear that the adult should endorse this utterance if he, the adult, feels (his own? the child's?) pain. Of course we know that this is not the case. Rather the adult will endorse the child's avowal if the child's behavior (crying, agitated motion, etc.) and, perhaps, the

[79] One detailed attempt to enter into such issues is Rogers Albritton, "On Wittgenstein's Use of the Term 'Criterion'," in Pitcher (ed.), *Wittgenstein: The Philosophical Investigations*, pp. 231–50, reprinted with a new postscript from *The Journal of Philosophy*, vol. 56 (1959), pp. 845–57.

[80] Following recent (perhaps not wholly attractive) philosophical usage, I call a first person assertion that the speaker has a certain sensation (e.g. "I am in pain") an 'avowal'.

external circumstances surrounding the child, indicate that he is in pain. If a child generally avows pain under such appropriate behavioral and external circumstances and generally does not do so otherwise, the adult will say of him that he has mastered the avowal, "I am in pain."

Since, in the case of discourse on pain and other sensations, the adult's confirmation whether he agrees with the child's avowal is based on the adult's observation of the child's behavior and circumstances, the fact that such behavior and circumstances characteristic of pain exist is essential in this case to the working of Wittgenstein's sceptical solution. This, then, is what is meant by the remark, "An 'inner process' stands in need of outward criteria." Roughly speaking, outward criteria for an inner process are circumstances, observable in the behavior of an individual, which, when present, would lead others to agree with his avowals. If the individual generally makes his avowals under the right such circumstances, others will say of him that he has mastered the appropriate expression ("I am in pain," "I feel itchy," etc.). We have seen that it is part of Wittgenstein's *general* view of the workings of *all* our expressions attributing concepts that others can confirm whether a subject's responses agree with their own. The present considerations simply spell out the form this confirmation and agreement take in the case of avowals.

It should then be clear that the demand for 'outward criteria' is no verificationist or behaviorist *premise* that Wittgenstein takes for granted in his 'private language argument'. If anything, it is *deduced*, in a sense of deduction akin to Kant's.[81]

[81] See also the postscript below, note 5.

Note that it would be difficult to imagine how a causal neurophysiological explanation of the uniformities in our attributions of sensations to others (of the type mentioned on p. 97 above) could be possible if there were no 'outward' manifestations of sensations. For – except perhaps in minute or subliminal ways – the sensations of one person are causally connected to those of others only by the mediation of external signs and behavior. (I assume that 'extrasensory perception' is not in question here.) If the mediating external correlates did not exist, how could the fact that others agree in their judgement that a given individual has a certain sensation have a causal explanation? Causally, it would have to be a coincidence. (Similarly for the uniformities in our mathematical judgements mentioned on pp. 105–6 below.)

However, Wittgenstein does not himself seem to be particularly concerned with neurophysiological explanations of such uniformities but wants to take them as 'protophenomena' (§§654–5), where the search for an explanation is a mistake. Although I do not think such remarks are meant to rule out causal neurophysiological explanations of the uniformities, it does not appear, philosophically, that Wittgenstein wishes to *rely* on the concept of such neurophysiological explanations either.

Obviously it *would* be incompatible with Wittgenstein's argument to seek to 'explain' our agreement on whether a given individual is in pain in terms of our uniform 'grasp' of the concept of *pain behavior*. The fact that we agree on whether a given individual is, or is not, say, groaning, comes within the purview of Wittgenstein's sceptical arguments as much as does any other case of 'following a rule'. The causal argument sketched above is something else. (Although I have tried to avoid invoking such an argument explicitly in my discussion of 'outward criteria' in the text, since – as I said – Wittgenstein does not seem to wish to rely on such considerations, it has sometimes seemed to me that such a causal argument is implicitly involved if it is to be argued that the criteria we actually use are *essential* to our 'language game' of attributing sensations.)

My discussion in this footnote and the preceding text was influenced by a question of G. E. M. Anscombe.

A sceptical problem is posed, and a sceptical solution to that problem is given. The solution turns on the idea that each person who claims to be following a rule can be checked by others. Others in the community can check whether the putative rule follower is or is not giving particular responses that they endorse, that agree with their own. The way they check this is, in general, a primitive part of the language game;[82] it need not operate the way it does in the case of 'table'.

[82] The criterion by which others judge whether a person is obeying a rule in a given instance cannot simply be his sincere inclination to say that he is, otherwise there would be no distinction between his thinking he is obeying the rule and his really obeying it (§202), and whatever he thinks is right will be right (§258). However, *after* the community judges (based on the original criteria) that he has mastered the appropriate rule, the community may (for certain rules) take the subject's sincere claim to follow it in this instance as in itself a new criterion for the correctness of

'Outward criteria' for sensations such as pain are simply the way this general requirement of our game of attributing concepts to others works out in the special case of sensations.[83]

---

his claim, without applying the original criteria. According to Wittgenstein, we do this in the case of 'I am in pain.' In the case of 'I dreamt', the terminology is originally taught to a subject who wakes up reporting certain experiences. We judge that he has mastered the rule for 'I dreamt' if he prefaces it to reports of experiences he says he had the night before. After we judge that he has mastered the language, we take 'I dreamt that such-and-such' as in itself a criterion for correctness. In both cases of 'I am in pain' and 'I dreamt', the first person utterance is new behavior that replaces the behavior that constituted the old criterion.

Reports of after-images or hallucinations are similar. We judge that someone has mastered 'I see something red' if he ordinarily utters it only when something red is present. Once we judge, however, that he has mastered this bit of language, we will accept his utterance that he sees red even when we think nothing red is present. Then we will say that he is suffering from an illusion, a hallucination, an after-image, or the like.

[83] One delicate point regarding sensations, and about 'criteria', ought to be noted. Wittgenstein often seems to be taken to suppose that for any type of sensation, there is an appropriate 'natural expression' of that sensation type ('pain behavior' for pain). The 'natural expression' is to be externally observable behavior 'expressing' the sensation other than, and prior to, the subject's verbal avowal that he has the sensation. If the theory of §244 that first person sensation avowals are verbal replacements for a 'primitive natural expression' of a sensation has the generality it appears to have, it would follow that Wittgenstein holds that such a 'primitive natural expression' must always exist if the first person avowal is to be meaningful. The impression is reinforced by other passages such as §§256–7. Further, the presentation of the private language argument in the present essay argues that for each rule I follow there must be a criterion – other than simply what I say – by which another will judge that I am following the rule correctly. Applied to sensations, this seems to mean that there must be some 'natural expression', or at any rate some external circumstances other than my mere inclination to say that this is the same sensation again, in virtue of which someone else can judge whether the sensation is present, and hence whether I have mastered the sensation term correctly. So the picture would be that to each statement of the form "I have sensation *S*" there must be an 'outward criterion' associated with *S*, other than the mere avowal itself, by which others recognize the presence or absence of *S*.

Not only professed followers of Wittgenstein but many who think of themselves as opponents (or, at least, not followers) of Wittgenstein, seem to think that something of this kind is true. That is to say, many philosophical programs seem to suppose that all sensation types are associated with some characteristic external phenomena (behavior, causes). In this essay I have largely suppressed my own views, which are by no means always in agreement with Wittgenstein's. However, I will permit myself to remark here that any view that supposes that, in this sense, an inner process always has 'outward criteria', seems to me probably to be *empirically* false. It seems to me that we have sensations or sensation *qualia* that we can perfectly well identify but that have no 'natural' external manifestations; an observer cannot tell in any way whether an individual has them unless that individual avows them. Perhaps a more liberal interpretation of the private language argument – which *may* be compatible with what Wittgenstein intended – would allow that a speaker might introduce some sensation terms with no 'outward criteria' for the associated sensations beyond his own sincere avowal of them. (Hence these avowals do not 'replace' any 'natural expressions' of the sensation(s), for there are none.) There will be no way anyone else will be in any position to check such a speaker, or to agree or disagree with him. (No matter what many Wittgensteinians – or Wittgenstein – would infer here, this does not in itself entail that his avowals are regarded as infallible, nor need it in itself mean that there could not later come to be ways of checking his avowals.) However, the language of the speaker, even his language of sensations, will not have the objectionable form of a 'private language', one in which anything he calls 'right' is right. The speaker can demonstrate, for many sensations that do have 'public criteria', that he has mastered the appropriate terminology for identifying these sensations. If we agree with his responses in enough cases of various sensations, we say of him that he has mastered 'sensation language'. All this, so far, is subject to external correction. But it is a primitive part of our language game of sensations that, if an individual has satisfied criteria for a mastery of sensation language in general, we then respect his claim to have identified a new type of sensation even if the sensation is correlated with nothing publicly observable. Then the only 'public criterion' for such an avowal will be the sincere avowal itself.

How does the view sketched here liberalize the private language argument as developed in the text? In the text we argued that *for each particular rule*, if conditionals of the form "If Jones follows the rule, in this instance he will . . ." are to have any point, they must be contraposed. If the community finds that in this instance Jones is not doing . . ., he is not following the rule. Only in this 'inverted' way does the notion of my behavior as 'guided' by the rule make sense. Thus for each rule there must be an 'external check' on whether I am following it in a given instance. Perhaps §202 should be taken to assert this. But this means the

It is not my purpose here to enter in detail into the exegesis of Wittgenstein's attack on an 'object and designation' model for sensation language (§293). I am not, in fact, sure that I fully

community must have a way of telling ('criterion') whether it is being followed in a given instance, which it uses to judge the speaker's mastery of the rule. This criterion cannot be simply the speaker's own sincere inclination to follow the rule a certain way – otherwise, the conditional has no content. This condition seems to be satisfied even in those cases where, *after* the community is satisfied that the speaker has mastered the language, it lets the speaker's sincere utterance be a (or *the*) criterion for their correctness. (See note 82.) In contrast, the liberal version allows that once a speaker, judged by criteria for mastery of various rules, is accepted into the community, there should be some rules where there is no way for others to check his mastery, but where that mastery is simply presumed on the basis of his membership in the community. This is simply a primitive feature of the language game. Why should Wittgenstein not allow language games like this?

I regret that I have discussed this matter so briefly in a note. I had thought at one time to expound the 'liberal' view sketched here as the 'official' Wittgenstein doctrine, which would have facilitated an exposition at greater length in the text. Certainly it is the one Wittgenstein should have adopted in accordance with the slogan "Don't think, look!", and it really is compatible with his attack on private language. On writing the final version of this essay, however, I came to worry that passages such as §244 and §§256–7 are highly misleading unless Wittgenstein holds something stronger.

(After writing the preceding, I found that Malcolm, in his *Thought and Knowledge* (Cornell University Press, Ithaca and London, 1977, 218 pp.), writes (p. 101), "philosophers sometimes read Wittgenstein's insistence on there being a conceptual link between statements of sensation and the primitive, natural, expressions of sensation in human behavior, as implying that there is a natural nonverbal, behavioral counterpart of *every* statement of sensation. Wittgenstein did not mean this, and it is obviously not true." I agree that it is not true. I think it is not true even for simple avowals invoking what we might call 'names' of sensations. ("I have sensation S.") But – what is a separate question – did Wittgenstein mean this? It seems to me that even some of Malcolm's own previous expositions of Wittgenstein have given (unintentionally?) the impression that he did, at least for simple avowals invoking 'names of sensations'. I myself have vacillated on the question. Whether or not Wittgenstein meant this, I do think that the essence of his doctrines can be captured without commitment to such a strong claim.)

understand it. But it seems likely that it relates to one aspect of our present considerations. The model of the way agreement operates with respect to a word like 'table' (perhaps a paradigm of 'object and designation') is a very simple one: the child says "Table!" when he sees that a table is present and the adult agrees if he also sees that a table is present. It is tempting to suppose that this model ought to be a general one, and that if it does not apply to the case of 'pain' we must conclude that in some sense the adult can never really confirm the correctness of the child's use of "I am in pain." Wittgenstein's suggestion is that there cannot and need not be such a demand based on generalizing the use of 'table'. No *a priori* paradigm of the way concepts ought to be applied governs all forms of life, or even our own form of life. Our game of attributing concepts to others depends on agreement. It so happens that in the case of ascribing sensation language, this agreement operates in part through 'outward criteria' for first person avowals. No further 'justification' or 'explanation' for this procedure is required; this simply is *given* as how we achieve agreement here. The important role played in our lives by the practice of attributing sensation concepts to others is evident. If I attribute mastery of the term 'pain' to someone, his sincere utterance of "I am in pain," even without other signs of pain, is sufficient to induce me to feel pity for him, attempt to aid him, and the like (or, if I am a sadist, for the opposite); and similarly in other cases.

Compare the case of mathematics. Mathematical statements are generally not about palpable entities: if they are indeed to be regarded as about 'entities', these 'entities' are generally suprasensible, eternal objects. And often mathematical statements are about the infinite. Even such an elementary mathematical truth as that any two integers have a unique sum (perhaps implicitly accepted by everyone who has mastered the concept of addition, and in any case, explicitly accepted by people with elementary sophistication as a basic property of that concept) is an assertion about infinitely many instances. All the more so is this true of the 'commutative'

law, that $x+y=y+x$ for all x and y. Yet how does agreement operate in the case of mathematics? How do we judge of someone else that he has mastered various mathematical concepts? Our judgement, as usual, stems from the fact that he agrees with us in enough particular cases of mathematical judgements (and that, even if he disagrees, we are operating with a common procedure). We do not compare his mind with some suprasensible, infinite reality: we have seen through the sceptical paradox that this is of no help if we ask, say, whether he has mastered the concept of addition. Rather we check his observable responses to particular addition problems to see if his responses agree with ours. In more sophisticated mathematical areas, he and we accept various mathematical statements on the basis of proof; and among the conditions we require for attributing to him the mastery of our mathematical concepts is his general agreement with us on what he regards as proof. Here 'proofs' are not abstract objects laid up in a mathematical heaven (say, lengthy proofs in a formal system such as *Principia*). They are visible (or audible or palpable), concrete phenomena – marks or diagrams on paper, intelligible utterances. Proofs in this sense are not only finite objects; they are also small and clear enough to be able to judge of another man's proof whether I too would regard it as proof. That is why Wittgenstein emphasizes that proof must be *surveyable*. It must be surveyable if it is to be usable as a basis for agreement in judgements.

This parallel illuminates Wittgenstein's remark that "Finitism and behaviorism are quite similar trends. Both say, but surely, all we have here is . . . Both deny the existence of something, both with a view to escaping from a confusion." (*Remarks on the Foundations of Mathematics*, p. 63 [II, §61]) How are the two trends 'quite similar'? The finitist realizes that although mathematical statements and concepts may be about the infinite (e.g. to grasp the '+' function is to grasp an infinite table), the criteria for attributing such functions to others must be 'finite', indeed 'surveyable' – for example, we attribute mastery of the concept of addition to a child on the basis of his

agreement with us on a finite number of instances of the addition table. Similarly, though sensation language may be about 'inner' states, the behaviorist correctly affirms that attribution to others of sensation concepts rests on publicly observable (and thus on behavioral) criteria. Further, the finitist and the behaviorist are right when they deny that the relation of the infinitary mathematical or inner psychological language to its 'finite' or 'outward' criteria is an adventitious product of human frailty, one that an account of the 'essence' of mathematical or sensation language would dispense with. Mathematical finitists and psychological behaviorists, however, make parallel unnecessary moves when they deny the legitimacy of talk of infinite mathematical objects or inner states. Behaviorists either condemn talk of mental states as meaningless or illegitimate, or attempt to define it in terms of behavior. Finitists similarly regard the infinitistic part of mathematics as meaningless. Such opinions are misguided: they are attempts to repudiate our ordinary language game. In this game we are allowed, for certain purposes, to assert statements about 'inner' states or mathematical functions under certain circumstances. Although the criteria for judging that such statements are legitimately introduced are indeed behavioral (or finite), finite or behavioral statements cannot replace their role in our language as we use it.

Let me, then, summarize the 'private language argument' as it is presented in this essay. (1) We all suppose that our language expresses concepts – 'pain', 'plus', 'red' – in such a way that, once I 'grasp' the concept, all future applications of it are determined (in the sense of being uniquely *justified* by the concept grasped). In fact, it seems that no matter what is in my mind at a given time, I am free in the future to interpret it in different ways – for example, I could follow the sceptic and interpret 'plus' as 'quus'. In particular, this point applies if I direct my attention to a sensation and name it; nothing I have done determines future applications (in the justificatory sense above). Wittgenstein's scepticism about the determination of future usage by the past contents of my mind is analogous to

Hume's scepticism about the determination of the future by the past (causally and inferentially). (2) The paradox can be resolved only by a 'sceptical solution of these doubts', in Hume's classic sense. This means that we must give up the attempt to find any fact about me in virtue of which I mean 'plus' rather than 'quus', and must then go on in a certain way. Instead we must consider how we actually use: (i) the categorical assertion that an individual is following a given rule (that he means addition by 'plus'); (ii) the conditional assertion that "if an individual follows such-and-such a rule, he must do so-and-so on a given occasion" (e.g., "if he means addition by '+', his answer to '68+57' should be '125'"). That is to say, we must look at the circumstances under which these assertions are introduced into discourse, and their role and utility in our lives. (3) As long as we consider a single individual in isolation, all we can say is this: An individual often does have the experience of being confident that he has 'got' a certain rule (sometimes that he has grasped it 'in a flash'). It is an empirical fact that, after that experience, individuals often are disposed to give responses in concrete cases with complete confidence that proceeding this way is 'what was intended'. We cannot, however, get any further in explaining on this basis the use of the conditionals in (ii) above. Of course, dispositionally speaking, the subject is indeed determined to respond in a certain way, say, to a given addition problem. Such a disposition, together with the appropriate 'feeling of confidence', could be present, however, even if he were not really following a rule at all, or even if he were doing the 'wrong' thing. The justificatory element of our use of conditionals such as (ii) is unexplained. (4) If we take into account the fact that the individual is in a community, the picture changes and the role of (i) and (ii) above becomes apparent. When the community accepts a particular conditional (ii), it accepts its *contraposed* form: the failure of an individual to come up with the particular responses the community regards as right leads the community to suppose that he is not following the rule. On the other hand, if an

individual passes enough tests, the community (endorsing assertions of the form (i)) accepts him as a rule follower, thus enabling him to engage in certain types of interactions with them that depend on their reliance on his responses. Note that this solution explains how the assertions in (i) and (ii) are introduced into language; it does *not* give conditions for these statements to be true. (5) The success of the practices in (3) depends on the brute empirical fact that we agree with each other in our responses. Given the sceptical argument in (1), this success cannot be explained by 'the fact that we all grasp the same concepts'. (6) Just as Hume thought he had demonstrated that the causal relation between two events is unintelligible unless they are subsumed under a regularity, so Wittgenstein thought that the considerations in (2) and (3) above showed that all talk of an individual following rules has reference to him as a member of a community, as in (3). In particular, for the conditionals of type (ii) to make sense, the community must be able to judge whether an individual is indeed following a given rule in particular applications, i.e. whether his responses agree with their own. In the case of avowals of sensations, the way the community makes this judgement is by observing the individual's behavior and surrounding circumstances.

A few concluding points regarding the argument ought to be noted. First, following §243, a 'private language' is usually defined as a language that is logically impossible for anyone else to understand. The private language argument is taken to argue against the possibility of a private language in this sense. This conception is not in error, but it seems to me that the emphasis is somewhat misplaced. What is really denied is what might be called the 'private model' of rule following, that the notion of a person following a given rule is to be analyzed simply in terms of facts about the rule follower and the rule follower alone, without reference to his membership in a wider community. (In the same way, what Hume denies is the private model of causation: that whether one event causes another is a matter of the relation between these two events

alone, without reference to their subsumption under larger event types.) The impossibility of a private language in the sense just defined does indeed follow from the incorrectness of the private model for language and rules, since the rule following in a 'private language' could only be analyzed by a private model, but the incorrectness of the private model is more basic, since it applies to all rules. I take all this to be the point of §202.

Does this mean that Robinson Crusoe, isolated on an island, cannot be said to follow any rules, no matter what he does?[84] I do not see that this follows. What does follow is that *if* we think of Crusoe as following rules, we are taking him into our community and applying our criteria for rule following to him.[85] The falsity of the private model need not mean that a *physically isolated* individual cannot be said to follow rules; rather that an individual, *considered in isolation* (whether or not he is physically isolated), cannot be said to do so. Remember that Wittgenstein's theory is one of assertability conditions. Our community can assert of any individual that he follows a rule if he passes the tests for rule following applied to any member of the community.

Finally, the point just made in the last paragraph, that

[84] See the well-known exchange between A. J. Ayer and Rush Rhees under the title "Can there be a Private Language?" (see note 47). Both participants in the exchange assume that the 'private language argument' excludes Crusoe from language. Ayer takes this alleged fact to be fatal to Wittgenstein's argument, while Rhees takes it to be fatal to Crusoe's language. Others, pointing out that a 'private language' is one that others *cannot* understand (see the preceding paragraph in the text), see no reason to think that the 'private language argument' has anything to do with Crusoe (as long as we could understand his language). My own view of the matter, as explained very briefly in the text, differs somewhat from all these opinions.

[85] If Wittgenstein would have any problem with Crusoe, perhaps the problem would be whether we have any 'right' to take him into our community in this way, and attribute our rules to him. See Wittgenstein's discussion of a somewhat similar question in §§199–200, and his conclusion, "Should we still be inclined to say they were playing a game? What right would one have to say so?"

Wittgenstein's theory is one of assertability conditions, deserves emphasis. Wittgenstein's theory should not be confused with a theory that, for any *m* and *n*, the value of the function we mean by 'plus', *is* (by definition) the value that (nearly) all the linguistic community would give as the answer. Such a theory would be a theory of the *truth* conditions of such assertions as "By 'plus' we mean such-and-such a function." or "By 'plus' we mean a function, which, when applied to 68 and 57 as arguments, yields 125 as value." (An infinite, exhaustive totality of specific conditions of the second form would determine which function was meant, and hence would determine a condition of the first form.) The theory would assert that 125 is the value of the function meant for given arguments, if and only if '125' is the response nearly everyone would give, given these arguments. Thus the theory would be a social, or community-wide, version of the dispositional theory, and would be open to at least some of the same criticisms as the original form. I take Wittgenstein to deny that he holds such a view, for example, in *Remarks on the Foundations of Mathematics*, V, §33 [VII, §40]: "Does this mean, e.g., that the definition of the same would be this: same is what all or most human beings take for the same? – Of course not."[86] (See also *Philosophical Investigations*, p. 226, "Certainly the propositions, "Human beings believe that twice two is four" and "Twice two is four" do not mean the same"; and see also §§240–1.) One must bear firmly in mind that Wittgenstein has no theory of truth conditions – necessary and sufficient conditions – for the correctness of one response rather than another to a new addition problem. Rather he simply points out that each of us *automatically* calculates new addition problems (without feeling the need to check with the community whether our procedure is proper); that the community feels entitled to correct a deviant calculation; that

[86] Although, in the passage in question, Wittgenstein is speaking of a particular language game of bringing something else and bringing the same, it is clear in context that it is meant to illustrate his general problem about rules. The entire passage is worth reading for the present issue.

in practice such deviation is rare, and so on. Wittgenstein thinks that these observations about sufficent conditions for justified assertion are enough to illuminate the role and utility in our lives of assertion about meaning and determination of new answers. What follows from these assertability conditions is *not* that the answer everyone gives to an addition problem is, by definition, the correct one, but rather the platitude that, if everyone agrees upon a certain answer, then no one will feel justified in calling the answer wrong.[87]

Obviously there are countless relevant aspects of Wittgenstein's philosophy of mind that I have not discussed. About some aspects I am not clear, and others have been left untouched because of the limits of this essay.[88] In particular, I have not discussed numerous issues arising out of the paragraphs *following* §243 that are usually called the 'private language argument', nor have I really discussed Wittgenstein's attendant positive account of the nature of sensation language and of the attribution of psychological states. Nevertheless, I do

[87] I feel some uneasiness may remain here. Considerations of time and space, as well as the fact that I might have to abandon the role of advocate in favor of that of critic, have prevented me from carrying out a more extensive discussion of this point.

[88] I might mention that, in addition to the Humean analogy emphasized in this essay, it has struck me that there is perhaps a certain analogy between Wittgenstein's private language argument and Ludwig von Mises's celebrated argument concerning economic calculation under socialism. (See e.g., his *Human Action* (2nd ed., Yale University Press, New Haven, 1963 xix+907 pp.), chapter 26, pp. 698–715, for one statement.) According to Mises, a rational economic calculator (say, the manager of an industrial plant) who wishes to choose the most efficient means to achieve given ends must compare alternative courses of action for cost effectiveness. To do this, he needs an array of prices (e.g. of raw materials, or machinery) set by *others*. If *one* agency set *all* prices, it could have no rational basis to choose between alternative courses of action. (Whatever seemed to it to be right would be right, so one cannot talk about right.) I do not know whether the fact bodes at all ill for the private language argument, but my impression is that although it is usually acknowledged that Mises's argument points to a real difficulty for centrally planned economies, it is now almost universally rejected as a theoretical proposition.

think that the basic 'private language argument' precedes these passages, and that only with an understanding of this argument can we begin to comprehend or consider what follows. That was the task undertaken in this essay.

WITTGENSTEIN, LUDWIG

PHILOSOPHISCHE UNTER-
SUCHUNGEN

TEIL I §§ 1-32

1. *Augustinus*, in den Confessionen 1/8: cum ipsi (majores homines) appellabant rem aliquam, et cum secundum eam vocem corpus ad aliquid movebant, videbam, et tenebam hoc ab eis vocari rem illam, quod sonabant, cum eam vellent ostendere. Hoc autem eos velle ex motu corporis aperiebatur: tamquam verbis naturalibus omnium gentium, quae fiunt vultu et nutu oculorum, ceterorumque membrorum actu, et sonitu vocis indicante affectionem animi in petendis, habendis, rejiciendis, fugiendisve rebus. Ita verba in variis sententiis locis suis posita, et crebro audita, quarum rerum signa essent, paulatim colligebam, measque jam voluntates, edomito in eis signis ore, per haec enuntiabam.

[Nannten die Erwachsenen irgend einen Gegenstand und wandten sie sich dabei ihm zu, so nahm ich das wahr und ich begriff, daß der Gegenstand durch die Laute, die sie aussprachen, bezeichnet wurde, da sie auf *ihn* hinweisen wollten. Dies aber entnahm ich aus ihren Gebärden, der natürlichen Sprache aller Völker, der Sprache, die durch Mienen- und Augenspiel, durch die Bewegungen der Glieder und den Klang der Stimme die Empfindungen der Seele anzeigt, wenn diese irgend etwas begehrt, oder festhält, oder zurückweist, oder flieht. So lernte ich nach und nach verstehen, welche Dinge die Wörter bezeichneten, die ich wieder und wieder, an ihren bestimmten Stellen in verschiedenen Sätzen, aussprechen hörte. Und ich brachte, als nun mein Mund sich an diese Zeichen gewöhnt hatte, durch sie meine Wünsche zum Ausdruck.]

In diesen Worten erhalten wir, so scheint es mir, ein bestimmtes Bild von dem Wesen der menschlichen Sprache. Nämlich dieses: Die Wörter der Sprache benennen Gegenstände—Sätze sind Verbindungen von solchen Benennungen.——In diesem Bild von der Sprache finden wir die Wurzeln der Idee: Jedes Wort hat eine Bedeutung. Diese Bedeutung ist dem Wort zugeordnet. Sie ist der Gegenstand, für welchen das Wort steht.

Von einem Unterschied der Wortarten spricht Augustinus nicht. Wer das Lernen der Sprache so beschreibt, denkt, so möchte ich glauben, zunächst an Hauptwörter, wie »Tisch«, »Stuhl«, »Brot«, und die Namen von Personen, erst in zweiter Linie an die Namen gewisser Tätigkeiten und Eigenschaften, und an die übrigen Wortarten als etwas, was sich finden wird.

Denke nun an diese Verwendung der Sprache: Ich schicke jemand einkaufen. Ich gebe ihm einen Zettel, auf diesem stehen die Zeichen: »fünf rote Äpfel«. Er trägt den Zettel zum Kaufmann; der öffnet die Lade, auf welcher das Zeichen »Äpfel« steht; dann sucht er in einer Tabelle das Wort »rot« auf und findet ihm gegenüber ein Farbmuster; nun sagt er die Reihe der Grundzahlwörter—ich nehme an, er weiß sie auswendig—bis zum Worte »fünf« und bei jedem Zahlwort nimmt er einen Apfel aus der Lade, der die Farbe des Musters hat.——So, und ähnlich, operiert man mit Worten.——»Wie weiß er aber, wo und wie er das Wort ›rot‹ nachschlagen soll und was er mit dem Wort ›fünf‹ anzufangen hat?«——Nun, ich nehme an, er *handelt,* wie ich es beschrieben habe. Die Erklärungen haben irgendwo ein Ende.—Was ist aber die Bedeutung des Wortes »fünf«? —Von einer solchen war hier garnicht die Rede; nur davon, wie das Wort »fünf« gebraucht wird.

2. Jener philosophische Begriff der Bedeutung ist in einer primitiven Vorstellung von der Art und Weise, wie die Sprache funktioniert, zu Hause. Man kann aber auch sagen, es sei die Vorstellung einer primitiveren Sprache als der unsern.

Denken wir uns eine Sprache, für die die Beschreibung, wie Augustinus sie gegeben hat, stimmt: Die Sprache soll der Verständigung eines Bauenden A mit einem Gehilfen B dienen. A führt einen Bau auf aus Bausteinen; es sind Würfel, Säulen, Platten und Balken vorhanden. B hat ihm die Bausteine zuzureichen, und zwar nach der Reihe, wie A sie braucht. Zu dem Zweck bedienen sie sich einer Sprache, bestehend aus den Wörtern: »Würfel«, »Säule«, »Platte«, »Balken«. A ruft sie aus;—B bringt den Stein, den er gelernt hat, auf diesen Ruf zu bringen.—— Fasse dies als vollständige primitive Sprache auf.

3. Augustinus beschreibt, könnten wir sagen, ein System der Verständigung; nur ist nicht alles, was wir Sprache nennen,

dieses System. Und das muß man in so manchen Fällen sagen, wo sich die Frage erhebt: »Ist diese Darstellung brauchbar, oder unbrauchbar?« Die Antwort ist dann: »Ja, brauchbar; aber nur für dieses eng umschriebene Gebiet, nicht für das Ganze, das du darzustellen vorgabst.«

Es ist, als erklärte jemand: »Spielen besteht darin, daß man Dinge, gewissen Regeln gemäß, auf einer Fläche verschiebt . . .« —und wir ihm antworten: Du scheinst an die Brettspiele zu denken; aber das sind nicht alle Spiele. Du kannst deine Erklärung richtigstellen, indem du sie ausdrücklich auf diese Spiele einschränkst.

4. Denk dir eine Schrift, in welcher Buchstaben zur Bezeichnung von Lauten benützt würden, aber auch zur Bezeichnung der Betonung und als Interpunktionszeichen. (Eine Schrift kann man auffassen als eine Sprache zur Beschreibung von Lautbildern.) Denk dir nun, daß Einer jene Schrift so verstünde, als entspräche einfach jedem Buchstaben ein Laut und als hätten die Buchstaben nicht auch ganz andere Funktionen. So einer, zu einfachen, Auffassung der Schrift gleicht Augustinus' Auffassung der Sprache.

5. Wenn man das Beispiel im §1 betrachtet, so ahnt man vielleicht, inwiefern der allgemeine Begriff der Bedeutung der Worte das Funktionieren der Sprache mit einem Dunst umgibt, der das klare Sehen unmöglich macht.—Es zerstreut den Nebel, wenn wir die Erscheinungen der Sprache an primitiven Arten ihrer Verwendung studieren, in denen man den Zweck und das Funktionieren der Wörter klar übersehen kann.

Solche primitiven Formen der Sprache verwendet das Kind, wenn es sprechen lernt. Das Lehren der Sprache ist hier kein Erklären, sondern ein Abrichten.

6. Wir könnten uns vorstellen, daß die Sprache im §2 die *ganze* Sprache des A und B ist; ja, die ganze Sprache eines Volksstamms. Die Kinder werden dazu erzogen, *diese* Tätigkeiten zu verrichten, *diese* Wörter dabei zu gebrauchen, und *so* auf die Worte des Anderen zu reagieren.

Ein wichtiger Teil der Abrichtung wird darin bestehen, daß der Lehrende auf die Gegenstände weist, die Aufmerksamkeit des Kindes auf sie lenkt, und dabei ein Wort ausspricht; z.B. das

Wort »Platte« beim Vorzeigen dieser Form. (Dies will ich nicht »hinweisende Erklärung«, oder »Definition«, nennen, weil ja das Kind noch nicht nach der Benennung *fragen* kann. Ich will es »hinweisendes Lehren der Wörter« nennen.——Ich sage, es wird einen wichtigen Teil der Abrichtung bilden, weil es bei Menschen so der Fall ist; nicht, weil es sich nicht anders vorstellen ließe.) Dieses hinweisende Lehren der Wörter, kann man sagen, schlägt eine assoziative Verbindung zwischen dem Wort und dem Ding: Aber was heißt das? Nun, es kann Verschiedenes heißen; aber man denkt wohl zunächst daran, daß dem Kind das Bild des Dings vor die Seele tritt, wenn es das Wort hört. Aber wenn das nun geschieht,—ist das der Zweck des Worts?—Ja, es *kann* der Zweck sein.—Ich kann mir eine solche Verwendung von Wörtern (Lautreihen) denken. (Das Aussprechen eines Wortes ist gleichsam ein Anschlagen einer Taste auf dem Vorstellungsklavier.) Aber in der Sprache im §2 ist es *nicht* der Zweck der Wörter, Vorstellungen zu erwecken. (Es kann freilich auch gefunden werden, daß dies dem eigentlichen Zweck förderlich ist.)

Wenn aber das das hinweisende Lehren bewirkt,—soll ich sagen, es bewirkt das Verstehen des Worts? Versteht nicht der den Ruf »Platte!«, der so und so nach ihm handelt?—Aber dies half wohl das hinweisende Lehren herbeiführen; aber doch nur zusammen mit einem bestimmten Unterricht. Mit einem anderen Unterricht hätte dasselbe hinweisende Lehren dieser Wörter ein ganz anderes Verständnis bewirkt.

»Indem ich die Stange mit dem Hebel verbinde, setze ich die Bremse instand.«—Ja, gegeben den ganzen übrigen Mechanismus. Nur mit diesem ist er der Bremshebel; und losgelöst von seiner Unterstützung ist er nicht einmal Hebel, sondern kann alles Mögliche sein, oder nichts.

7. In der Praxis des Gebrauchs der Sprache (2) ruft der eine Teil die Wörter, der andere handelt nach ihnen; im Unterricht der Sprache aber wird sich *dieser* Vorgang finden: Der Lernende *benennt* die Gegenstände. D.h. er spricht das Wort, wenn der Lehrer auf den Stein zeigt.—Ja, es wird sich hier die noch einfachere Übung finden: der Schüler spricht die Worte nach, die der Lehrer ihm vorsagt—beides sprachähnliche Vorgänge.

Wir können uns auch denken, daß der ganze Vorgang des

Gebrauchs der Worte in (2) eines jener Spiele ist, mittels welcher Kinder ihre Muttersprache erlernen. Ich will diese Spiele *»Sprachspiele«* nennen, und von einer primitiven Sprache manchmal als einem Sprachspiel reden.

Und man könnte die Vorgänge des Benennens der Steine und des Nachsprechens des vorgesagten Wortes auch Sprachspiele nennen. Denke an manchen Gebrauch, der von Worten in Reigenspielen gemacht wird.

Ich werde auch das Ganze: der Sprache und der Tätigkeiten, mit denen sie verwoben ist, das »Sprachspiel« nennen.

8. Sehen wir eine Erweiterung der Sprache (2) an. Außer den vier Wörtern »Würfel«, »Säule«, etc. enthalte sie eine Wörterreihe, die verwendet wird, wie der Kaufmann in (1) die Zahlwörter verwendet (es kann die Reihe der Buchstaben des Alphabets sein); ferner, zwei Wörter, sie mögen »dorthin« und »dieses« lauten (weil dies schon ungefähr ihren Zweck andeutet), sie werden in Verbindung mit einer zeigenden Handbewegung gebraucht; und endlich eine Anzahl von Farbmustern. A gibt einen Befehl von der Art: »d-Platte-dorthin«. Dabei läßt er den Gehilfen ein Farbmuster sehen, und beim Worte »dorthin« zeigt er an eine Stelle des Bauplatzes. B nimmt von dem Vorrat der Platten je eine von der Farbe des Musters für jeden Buchstaben des Alphabets bis zum »d« und bringt sie an den Ort, den A bezeichnet.—Bei anderen Gelegenheiten gibt A den Befehl: »dieses-dorthin«. Bei »dieses« zeigt er auf einen Baustein. Usw.

9. Wenn das Kind diese Sprache lernt, muß es die Reihe der ›Zahlwörter‹ a, b, c, . . . auswendiglernen. Und es muß ihren Gebrauch lernen.—Wird in diesem Unterricht auch ein hinweisendes Lehren der Wörter vorkommen?—Nun, es wird z.B. auf Platten gewiesen und gezählt werden: »a, b, c Platten«.—Mehr Ähnlichkeit mit dem hinweisenden Lehren der Wörter »Würfel«, »Säule«, etc. hätte das hinweisende Lehren von Zahlwörtern, die nicht zum Zählen dienen, sondern zur Bezeichnung mit dem Auge erfaßbarer Gruppen von Dingen. So lernen ja Kinder den Gebrauch der ersten fünf oder sechs Grundzahlwörter.

Wird auch »dorthin« und »dieses« hinweisend gelehrt?—Stell dir vor, wie man ihren Gebrauch etwa lehren könnte! Es wird dabei auf Örter und Dinge gezeigt werden,—aber hier geschieht

ja dieses Zeigen auch im *Gebrauch* der Wörter und nicht nur beim Lernen des Gebrauchs.—

10. Was *bezeichnen* nun die Wörter dieser Sprache?—Was sie bezeichnen, wie soll ich das zeigen, es sei denn in der Art ihres Gebrauchs? Und den haben wir ja beschrieben. Der Ausdruck »dieses Wort bezeichnet *das*« müßte also ein Teil dieser Beschreibung werden. Oder: die Beschreibung soll auf die Form gebracht werden. »Das Wort .... bezeichnet ....«.

Nun, man kann ja die Beschreibung des Gebrauchs des Wortes »Platte« dahin abkürzen, daß man sagt, dieses Wort bezeichne diesen Gegenstand. Das wird man tun, wenn es sich z.B. nurmehr darum handelt, das Mißverständnis zu beseitigen, das Wort »Platte« beziehe sich auf die Bausteinform, die wir tatsächlich »Würfel« nennen,— die Art und Weise dieses ›*Bezugs*‹ aber, d.h. der Gebrauch dieser Worte im übrigen, bekannt ist.

Und ebenso kann man sagen, die Zeichen »a«, »b«, etc. bezeichnen Zahlen; wenn dies etwa das Mißverständnis behebt; »a«, »b«, »c«, spielten in der Sprache die Rolle, die in Wirklichkeit »Würfel, »Platte«, »Säule«, spielen. Und man kann auch sagen, »c« bezeichne diese Zahl und nicht jene; wenn damit etwa erklärt wird, die Buchstaben seien in der Reihenfolge a, b, c, d, etc. zu verwenden und nicht in der: a, b, d, c.

Aber dadurch, daß man so die Beschreibungen des Gebrauchs der Wörter einander anähnelt, kann doch dieser Gebrauch nicht ähnlicher werden! Denn, wie wir sehen, ist er ganz und gar ungleichartig.

11. Denk an die Werkzeuge in einem Werkzeugkasten: es ist da ein Hammer, eine Zange, eine Säge, ein Schraubenzieher, ein Maßstab, ein Leimtopf, Leim, Nägel und Schrauben.—So verschieden die Funktionen dieser Gegenstände, so verschieden sind die Funktionen der Wörter. (Und es gibt Ähnlichkeiten hier und dort.)

Freilich, was uns verwirrt ist die Gleichförmigkeit ihrer Erscheinung, wenn die Wörter uns gesprochen, oder in der Schrift und im Druck entgegentreten. Denn ihre *Verwendung* steht nicht so deutlich vor uns. Besonders nicht, wenn wir philosophieren!

12. Wie wenn wir in den Führerstand einer Lokomotive schauen: da sind Handgriffe, die alle mehr oder weniger gleich

aussehen. (Das ist begreiflich, denn sie sollen alle mit der Hand angefaßt werden.) Aber einer ist der Handgriff einer Kurbel, die kontinuierlich verstellt werden kann (sie reguliert die Öffnung eines Ventils); ein andrer ist der Handgriff eines Schalters, der nur zweierlei wirksame Stellungen hat, er ist entweder umgelegt, oder aufgestellt; ein dritter ist der Griff eines Bremshebels, je stärker man zieht, desto stärker wird gebremst; ein vierter, der Handgriff einer Pumpe, er wirkt nur, solange er hin und her bewegt wird.

13. Wenn wir sagen: »jedes Wort der Sprache bezeichnet etwas« so ist damit vorerst noch *gar* nichts gesagt; es sei denn, daß wir genau erklärten, *welche* Unterscheidung wir zu machen wünschen. (Es könnte ja sein, daß wir die Wörter der Sprache (8) von Wörtern ›ohne Bedeutung‹ unterscheiden wollten, wie sie in Gedichten Lewis Carroll's vorkommen, oder von Worten wie »juwiwallera« in einem Lied.)

14. Denke dir, jemand sagte: »*Alle* Werkzeuge dienen dazu, etwas zu modifizieren. So, der Hammer die Lage des Nagels, die Säge die Form des Bretts, etc.«—Und was modifiziert der Maßstab, der Leimtopf, die Nägel?—»Unser Wissen um die Länge eines Dings, die Temperatur des Leims, und die Festigkeit der Kiste.«——Wäre mit dieser Assimilation des Ausdrucks etwas gewonnen?—

15. Am direktesten ist das Wort »bezeichnen« vielleicht da angewandt, wo das Zeichen auf dem Gegenstand steht, den es bezeichnet. Nimm an, die Werkzeuge, die A beim Bauen benützt, tragen gewisse Zeichen. Zeigt A dem Gehilfen ein solches Zeichen, so bringt dieser das Werkzeug, das mit dem Zeichen versehen ist.

So, und auf mehr oder weniger ähnliche Weise, bezeichnet ein Name ein Ding, und wird ein Name einem Ding gegeben.—Es wird sich oft nützlich erweisen, wenn wir uns beim Philosophieren sagen: Etwas benennen, das ist etwas Ähnliches, wie einem Ding ein Namentäfelchen anheften.

16. Wie ist es mit den Farbmustern, die A dem B zeigt,—gehören sie zur *Sprache*? Nun, wie man will. Zur Wortsprache gehören sie nicht; aber wenn ich jemandem sage: »Sprich das Wort ›das‹ aus«, so wirst du doch dieses zweite »›das‹« auch noch zum

Satz rechnen. Und doch spielt es eine ganz ähnliche Rolle, wie ein Farbmuster im Sprachspiel (8); es ist nämlich ein Muster dessen, was der Andre sagen soll.

Es ist das Natürlichste, und richtet am wenigsten Verwirrung an, wenn wir die Muster zu den Werkzeugen der Sprache rechnen.

((Bemerkung über das reflexive Fürwort *»dieser* Satz«.))

17. Wir werden sagen können: in der Sprache (8) haben wir verschiedene *Wortarten*. Denn die Funktion des Wortes »Platte« und des Wortes »Würfel« sind einander ähnlicher als die von »Platte« und von »d«. Wie wir aber die Worte nach Arten zusammenfassen, wird vom Zweck der Einteilung abhängen,—und von unserer Neigung.

Denke an die verschiedenen Gesichtspunkte, nach denen man Werkzeuge in Werkzeugarten einteilen kann. Oder Schachfiguren in Figurenarten.

18. Daß die Sprachen (2) und (8) nur aus Befehlen bestehen, laß dich nicht stören. Willst du sagen, sie seien darum nicht vollständig, so frage dich, ob unsere Sprache vollständig ist;—ob sie es war, ehe ihr der chemische Symbolismus und die Infinitesimalnotation einverleibt wurden; denn dies sind, sozusagen, Vorstädte unserer Sprache. (Und mit wieviel Häusern, oder Straßen, fängt eine Stadt an, Stadt zu sein?) Unsere Sprache kann man ansehen als eine alte Stadt: Ein Gewinkel von Gäßchen und Plätzen, alten und neuen Häusern, und Häusern mit Zubauten aus verschiedenen Zeiten; und dies umgeben von einer Menge neuer Vororte mit geraden und regelmäßigen Straßen und mit einförmigen Häusern.

19. Man kann sich leicht eine Sprache vorstellen, die nur aus Befehlen und Meldungen in der Schlacht besteht.—Oder eine Sprache, die nur aus Fragen besteht und einem Ausdruck der Bejahung und der Verneinung. Und unzählige Andere.——Und eine Sprache vorstellen heißt, sich eine Lebensform vorstellen.

Wie ist es aber: Ist der Ruf »Platte!« im Beispiel (2) ein Satz oder ein Wort?—Wenn ein Wort, so hat es doch nicht dieselbe Bedeutung wie das gleichlautende unserer gewöhnlichen Sprache, denn im §2 ist es ja ein Ruf. Wenn aber ein Satz, so ist es doch nicht der elliptische Satz »Platte!« unserer Sprache. ——Was die erste Frage anbelangt, so kannst du »Platte!« ein Wort, und

auch einen Satz nennen; vielleicht treffend einen ›degenerierten Satz‹ (wie man von einer degenerierten Hyperbel spricht), und zwar ist es eben unser ›elliptischer‹ Satz.—Aber der ist doch nur eine verkürzte Form des Satzes »Bring mir eine Platte!« und diesen Satz gibt es doch in Beispiel (2) nicht.—Aber warum sollte ich nicht, umgekehrt, den Satz »Bring mir eine Platte!« eine *Verlängerung* des Satzes »Platte!« nennen?—Weil der, der »Platte!« ruft, eigentlich meint: »Bring mir eine Platte!«.—Aber wie machst du das, *dies meinen*, während du »Platte« *sagst*? Sprichst du dir inwendig den unverkürzten Satz vor? Und warum soll ich, um zu sagen, was Einer mit dem Ruf »Platte« meint, diesen Ausdruck in einen andern übersetzen? Und wenn sie das Gleiche bedeuten,—warum soll ich nicht sagen: »wenn er ›Platte!‹ sagt, meint er ›Platte!‹«? Oder: warum sollst du nicht »Platte« meinen können, wenn du »Bring mir die Platte« meinen kannst?——Aber wenn ich »Platte!« rufe, so will ich doch, *er soll mir eine Platte bringen!*——Gewiß, aber besteht ›dies wollen‹ darin, daß du in irgend einer Form einen andern Satz denkst als den, den du sagst?—

20. Aber wenn nun Einer sagt »Bring mir eine Platte!«, so scheint es ja jetzt, als könnte er diesen Ausdruck als *ein* langes Wort meinen: entsprechend nämlich dem einen Worte »Platte!«. ——Kann man ihn also einmal als *ein* Wort, einmal als vier Wörter meinen? Und wie meint man ihn gewöhnlich?——Ich glaube, wir werden geneigt sein, zu sagen: Wir meinen den Satz als einen von *vier* Wörtern, wenn wir ihn im Gegensatz zu andern Sätzen gebrauchen, wie »*Reich* mir eine Platte zu«, »Bring *ihm* eine Platte«, »Bring *zwei* Platten«,etc.; also im Gegensatz zu Sätzen, welche die Wörter unseres Befehls in andern Verbindungen enthalten.——Aber worin besteht es, einen Satz im Gegensatz zu andern Sätzen gebrauchen? Schweben einem dabei etwa diese Sätze vor? Und *alle*? Und *während* man den einen Satz sagt, oder vor-, oder nachher?—Nein! Wenn auch so eine Erklärung einige Versuchung für uns hat, so brauchen wir doch nur einen Augenblick zu bedenken, was vielleicht geschieht, um zu sehen, daß wir hier auf falschem Weg sind. Wir sagen, wir gebrauchen den Befehl im Gegensatz zu andern Sätzen, weil *unsere Sprache* die Möglichkeit dieser andern Sätze enthält. Wer

unsere Sprache nicht versteht, ein Ausländer, der öfter gehört hätte, wie jemand den Befehl gibt »Bring mir eine Platte!«, könnte der Meinung sein, diese ganze Lautreihe sei ein Wort und entspräche etwa dem Wort für »Baustein« in seiner Sprache. Wenn er selbst dann diesen Befehl gegeben hätte, würde er ihn vielleicht anders aussprechen, und wir würden sagen: Er spricht ihn so sonderbar aus, weil er ihn für *ein* Wort hält.—— Aber geht also nicht, wenn er ihn ausspricht, eben auch etwas anderes in ihm vor,—*dem* entsprechend, daß er den Satz als *ein* Wort auffaßt.—Es kann das Gleiche in ihm vorgehen, oder auch anderes. Was geht denn in dir vor, wenn du so einen Befehl gibst; bist du dir bewußt, daß er aus vier Wörtern besteht, *während* du ihn aussprichst? Freilich, du *beherrschst* diese Sprache—in der es auch jene andern Sätze gibt—aber ist dieses Beherrschen etwas, was ›*geschieht*‹, während du den Satz aussprichst?—Und ich habe ja zugegeben: der Fremde wird den Satz, den er anders auffaßt, wahrscheinlich anders aussprechen; aber, was wir die falsche Auffassung nennen, *muß* nicht in irgend etwas liegen, was das Aussprechen des Befehls begleitet.

›Elliptisch‹ ist der Satz nicht, weil er etwas ausläßt, was wir meinen, wenn wir ihn aussprechen, sondern weil er gekürzt ist—im Vergleich mit einem bestimmten Vorbild unserer Grammatik.—Man könnte hier freilich den Einwand machen: »Du gibst zu, daß der verkürzte und der unverkürzte Satz den gleichen Sinn haben.—Welchen Sinn haben sie also? Gibt es denn für diesen Sinn nicht einen Wortausdruck?«——Aber besteht der gleiche Sinn der Sätze nicht in ihrer gleichen *Verwendung*?—(Im Russischen heißt es »Stein rot« statt »der Stein ist rot«; geht ihnen die Kopula im Sinn ab, oder *denken* sie sich die Kopula dazu?)

21. Denke dir ein Sprachspiel, in welchem B dem A auf dessen Frage die Anzahl der Platten, oder Würfel in einem Stoß meldet, oder die Farben und Formen der Bausteine, die dort und dort liegen.—So eine Meldung könnte also lauten: »Fünf Platten«. Was ist nun der Unterschied zwischen der Meldung, oder Behauptung, »Fünf Platten« und dem Befehl »Fünf Platten!«?—Nun, die Rolle, die das Aussprechen dieser Worte im Sprachspiel spielt. Aber es wird wohl auch der Ton, in dem sie ausgesprochen

werden, ein anderer sein, und die Miene, und noch manches andere. Aber wir können uns auch denken, daß der Ton der gleiche ist,—denn ein Befehl und eine Meldung können in *mancherlei* Ton ausgesprochen werden und mit mancherlei Miene—und daß der Unterschied allein in der Verwendung liegt. (Freilich könnten wir auch die Worte »Behauptung« und »Befehl« zur Bezeichnung einer grammatischen Satzform und eines Tonfalls gebrauchen; wie wir ja »Ist das Wetter heute nicht herrlich?« eine Frage nennen, obwohl sie als Behauptung verwendet wird.) Wir könnten uns eine Sprache denken, in der *alle* Behauptungen die Form und den Ton rhetorischer Fragen hätten; oder jeder Befehl die Form der Frage: »Möchtest du das tun?« Man wird dann vielleicht sagen: »Was er sagt, hat die Form der Frage, ist aber wirklich ein Befehl«—d.h., hat die Funktion des Befehls in der Praxis der Sprache. (Ähnlich sagt man »Du wirst das tun«, nicht als Prophezeihung, sondern als Befehl. Was macht es zu dem einen, was zu dem andern?)

22. Freges Ansicht, daß in einer Behauptung eine Annahme steckt, die dasjenige ist, was behauptet wird, basiert eigentlich auf der Möglichkeit, die es in unserer Sprache gibt, jeden Behauptungssatz in der Form zu schreiben »Es wird behauptet, daß das und das der Fall ist.«—Aber »Daß das und das der Fall ist« ist eben in unsrer Sprache kein Satz—es ist noch kein *Zug* im Sprachspiel. Und schreibe ich statt »Es wird behauptet, daß . . .« »Es wird behauptet: das und das ist der Fall«, dann sind hier die Worte »Es wird behauptet« eben überflüssig.

Wir könnten sehr gut auch jede Behauptung in der Form einer Frage mit nachgesetzter Bejahung schreiben; etwa: »Regnet es? Ja!« Würde das zeigen, daß in jeder Behauptung eine Frage steckt?

Denken wir uns ein Bild, einen Boxer in bestimmter Kampfstellung darstellend. Dieses Bild kann nun dazu gebraucht werden, um jemand mitzuteilen, wie er stehen, sich halten soll; oder, wie er sich nicht halten soll; oder, wie ein bestimmter Mann dort und dort gestanden hat; oder etc. etc. Man könnte dieses Bild (chemisch gesprochen) ein Satzradikal nennen. Ähnlich dachte sich wohl Frege die »Annahme«.

Man hat wohl das Recht, ein Behauptungszeichen zu verwenden im Gegensatz z.B. zu einem Fragezeichen; oder wenn man eine Behauptung unterscheiden will von einer Fiktion, oder einer Annahme. Irrig ist es nur, wenn man meint, daß die Behauptung nun aus zwei Akten besteht, dem Erwägen und dem Behaupten (Beilegen des Wahrheitswerts, oder dergl.) und daß wir diese Akte nach dem Zeichen des Satzes vollziehen, ungefähr wie wir nach Noten singen. Mit dem Singen nach Noten ist allerdings das laute, oder leise Lesen des geschriebenen Satzes zu vergleichen, aber nicht das ›*Meinen*‹ (Denken) des gelesenen Satzes.

Das Fregesche Behauptungszeichen betont den *Satzanfang*. Es hat also eine ähnliche Funktion wie der Schlußpunkt. Es unterscheidet die ganze Periode vom Satz *in* der Periode. Wenn ich Einen sagen höre »es regnet«, aber nicht weiß, ob ich den Anfang und den Schluß der Periode gehört habe, so ist dieser Satz für mich noch kein Mittel der Verständigung.

23. Wieviele Arten der Sätze gibt es aber? Etwa Behauptung, Frage und Befehl?—Es gibt *unzählige* solcher Arten: unzählige verschiedene Arten der Verwendung alles dessen, was wir »Zeichen«, »Worte«, »Sätze«, nennen. Und diese Mannigfaltigkeit ist nichts Festes, ein für allemal Gegebenes; sondern neue Typen der Sprache, neue Sprachspiele, wie wir sagen können, entstehen und andre veralten und werden vergessen. (Ein *ungefähres Bild* davon können uns die Wandlungen der Mathematik geben.)

Das Wort »Sprach*spiel*« soll hier hervorheben, daß das Sprechen der Sprache ein Teil ist einer Tätigkeit, oder einer Lebensform.

Führe dir die Mannigfaltigkeit der Sprachspiele an diesen Beispielen, und anderen, vor Augen:

Befehlen, und nach Befehlen handeln—
Beschreiben eines Gegenstands nach dem Ansehen, oder nach Messungen—
Herstellen eines Gegenstands nach einer Beschreibung (Zeichnung)—
Berichten eines Hergangs—
Über den Hergang Vermutungen anstellen—
Eine Hypothese aufstellen und prüfen—
Darstellen der Ergebnisse eines Experiments durch Tabellen und Diagramme—

Eine Geschichte erfinden; und lesen—
Theater spielen—
Reigen singen—
Rätsel raten—
Einen Witz machen; erzählen—
Ein angewandtes Rechenexempel lösen—
Aus einer Sprache in die andere übersetzen—
Bitten, Danken, Fluchen, Grüßen, Beten.

—Es ist interessant, die Mannigfaltigkeit der Werkzeuge der Sprache und ihrer Verwendungsweisen, die Mannigfaltigkeit der Wort- und Satzarten, mit dem zu vergleichen, was Logiker über den Bau der Sprache gesagt haben. (Und auch der Verfasser der *Logisch-Philosophischen Abhandlung*.)

24. Wem die Mannigfaltigkeit der Sprachspiele nicht vor Augen ist, der wird etwa zu den Fragen geneigt sein, wie dieser: »Was ist eine Frage?«—Ist es die Feststellung, daß ich das und das nicht weiß, oder die Feststellung, daß ich wünsche, der Andre möchte mir sagen....? Oder ist es die Beschreibung meines seelischen Zustandes der Ungewißheit?—Und ist der Ruf »Hilfe!« so eine Beschreibung?

Denke daran, wieviel Verschiedenartiges »Beschreibung« genannt wird: Beschreibung der Lage eines Körpers durch seine Koordinaten; Beschreibung eines Gesichtsausdrucks; Beschreibung einer Tastempfindung; einer Stimmung.

Man kann freilich statt der gewöhnlichen Form der Frage die der Feststellung, oder Beschreibung setzen: »Ich will wissen, ob....«, oder »Ich bin im Zweifel, ob....«—aber damit hat man die verschiedenen Sprachspiele einander nicht näher gebracht.

Die Bedeutsamkeit solcher Umformungsmöglichkeiten, z.B. aller Behauptungssätze in Sätze, die mit der Klausel »Ich denke«, oder »Ich glaube« anfangen (also sozusagen in Beschreibungen *meines* Innenlebens) wird sich an anderer Stelle deutlicher zeigen. (Solipsismus.)

25. Man sagt manchmal: die Tiere sprechen nicht, weil ihnen die geistigen Fähigkeiten fehlen. Und das heißt: »sie denken nicht, darum sprechen sie nicht«. Aber: sie sprechen eben nicht. Oder besser: sie verwenden die Sprache nicht—wenn wir von

den primitivsten Sprachformen absehen.—Befehlen, fragen, erzählen, plauschen gehören zu unserer Naturgeschichte so wie gehen, essen, trinken, spielen.

26. Man meint, das Lernen der Sprache bestehe darin, daß man Gegenstände benennt. Und zwar: Menschen, Formen, Farben, Schmerzen, Stimmungen, Zahlen etc. Wie gesagt—das Benennen ist etwas Ähnliches, wie, einem Ding ein Namentäfelchen anheften. Man kann das eine Vorbereitung zum Gebrauch eines Wortes nennen. Aber *worauf* ist es eine Vorbereitung?

27. »Wir benennen die Dinge und können nun über sie reden. Uns in der Rede auf sie beziehen.«—Als ob mit dem Akt des Benennens schon das, was wir weiter tun, gegeben wäre. Als ob es nur Eines gäbe, was heißt: »von Dingen reden«. Während wir doch das Verschiedenartigste mit unsern Sätzen tun. Denken wir allein an die Ausrufe. Mit ihren ganz verschiedenen Funktionen.

Wasser!
Fort!
Au!
Hilfe!
Schön!
Nicht!

Bist du nun noch geneigt, diese Wörter »Benennungen von Gegenständen« zu nennen?

In den Sprachen (2) und (8) gab es ein Fragen nach der Benennung nicht. Dies und sein Korrelat, die hinweisende Erklärung, ist, wie wir sagen könnten, ein eigenes Sprachspiel. Das heißt eigentlich: wir werden erzogen, abgerichtet dazu, zu fragen: »Wie heißt das?«—worauf dann das Benennen erfolgt. Und es gibt auch ein Sprachspiel: Für etwas einen Namen erfinden. Also, zu sagen: »Das heißt . . . .«, und nun den neuen Namen zu verwenden. (So benennen Kinder z.B. ihre Puppen und reden dann von ihnen, und zu ihnen. Dabei bedenke gleich, wie eigenartig der Gebrauch des Personennamens ist, mit welchem wir den Benannten *rufen*!)

28. Man kann nun einen Personennamen, ein Farbwort, einen Stoffnamen, ein Zahlwort, den Namen einer Himmelsrichtung, etc. hinweisend definieren. Die Definition der Zahl Zwei »Das heißt ›zwei‹«—wobei man auf zwei Nüsse zeigt—ist vollkommen exakt.—Aber wie kann man denn die Zwei so definieren? Der, dem man die Definition gibt, weiß ja dann nicht, *was* man mit »zwei« benennen will; er wird annehmen, daß du *diese* Gruppe von Nüssen »zwei« nennst!—Er *kann* dies annehmen; vielleicht nimmt er es aber nicht an. Er könnte ja auch, umgekehrt, wenn ich dieser Gruppe von Nüssen einen Namen beilegen will, ihn als Zahlnamen mißverstehen. Und ebensogut, wenn ich einen Personennamen hinweisend erkläre, diesen als Farbnamen, als Bezeichnung der Rasse, ja als Namen einer Himmelsrichtung auffassen. Das heißt, die hinweisende Definition kann in *jedem* Fall so und anders gedeutet werden.

29. Vielleicht sagt man: die Zwei kann nur *so* hinweisend definiert werden: »Diese *Zahl* heißt ›zwei‹«. Denn das Wort »Zahl« zeigt hier an, an welchen *Platz* der Sprache, der Grammatik, wir das Wort setzen. Das heißt aber, es muß das Wort »Zahl« erklärt sein, ehe jene hinweisende Definition verstanden werden kann.—Das Wort »Zahl« in der Definition zeigt allerdings diesen Platz an; den Posten, an den wir das Wort stellen. Und wir können so Mißverständnissen vorbeugen, indem wir sagen: »Diese *Farbe* heißt so und so«, »Diese *Länge* heißt so und so«, usw. Das heißt: Mißverständnisse werden manchmal so vermieden. Aber läßt sich denn das Wort »Farbe«, oder »Länge« nur so auffassen?—Nun, wir müssen sie eben erklären.—Also erklären durch andere Wörter! Und wie ist es mit der letzten Erklärung in dieser Kette? (Sag nicht »Es gibt keine ›letzte‹ Erklä-

---

Könnte man zur Erklärung des Wortes »rot« auf etwas weisen, was *nicht rot* ist? Das wäre ähnlich, wie wenn man Einem, der der deutschen Sprache nicht mächtig ist, das Wort »bescheiden« erklären sollte, und man zeigte zur Erklärung auf einen arroganten Menschen und sagte »Dieser ist *nicht* bescheiden«. Es ist kein Argument gegen eine solche Erklärungsweise, daß sie vieldeutig ist. Jede Erklärung kann mißverstanden werden.

Wohl aber könnte man fragen: Sollen wir das noch eine »Erklärung« nennen?—Denn sie spielt im Kalkül natürlich eine andere Rolle als das, was wir gewöhnlich »hinweisende Erklärung« des Wortes »rot« nennen; auch wenn sie dieselben praktischen Folgen, dieselbe *Wirkung* auf den Lernenden hätte.

rung«. Das ist gerade so, als wolltest du sagen: »Es gibt kein letztes Haus in dieser Straße; man kann immer noch eines dazubauen.«)

Ob das Wort »Zahl« in der hinweisenden Definition der Zwei nötig ist, das hängt davon ab, ob er sie ohne dieses Wort anders auffaßt, als ich es wünsche. Und das wird wohl von den Umständen abhängen, unter welchen sie gegeben wird, und von dem Menschen, dem ich sie gebe.

Und wie er die Erklärung ›auffaßt‹, zeigt sich darin, wie er von dem erklärten Wort Gebrauch macht.

30. Man könnte also sagen: Die hinweisende Definition erklärt den Gebrauch—die Bedeutung—des Wortes, wenn es schon klar ist, welche Rolle das Wort in der Sprache überhaupt spielen soll. Wenn ich also weiß, daß Einer mir ein Farbwort erklären will, so wird mir die hinweisende Erklärung »Das heißt ›Sepia‹« zum Verständnis des Wortes verhelfen.— Und dies kann man sagen, wenn man nicht vergißt, daß sich nun allerlei Fragen an das Wort »wissen«, oder »klar sein« anknüpfen.

Man muß schon etwas wissen (oder können), um nach der Benennung fragen zu können. Aber was muß man wissen?

31. Wenn man jemandem die Königsfigur im Schachspiel zeigt und sagt »Das ist der Schachkönig«, so erklärt man ihm dadurch nicht den Gebrauch dieser Figur,—es sei denn, daß er die Regeln des Spiels schon kennt, bis auf diese letzte Bestimmung: die Form einer Königsfigur. Man kann sich denken, er habe die Regeln des Spiels gelernt, ohne daß ihm je eine wirkliche Spielfigur gezeigt wurde. Die Form der Spielfigur entspricht hier dem Klang, oder der Gestalt eines Wortes.

Man kann sich aber auch denken, Einer habe das Spiel gelernt, ohne je Regeln zu lernen, oder zu formulieren. Er hat etwa zuerst durch Zusehen ganz einfache Brettspiele gelernt und ist zu immer komplizierteren fortgeschritten. Auch diesem könnte man die Erklärung geben: »Das ist der König«—wenn man ihm z.B. Schachfiguren von einer ihm ungewohnten Form zeigt. Auch diese Erklärung lehrt ihn den Gebrauch der Figur nur darum, weil, wie wir sagen könnten, der Platz schon vorbereitet war an den sie gestellt wurde. Oder auch: Wir werden nur dann sagen, sie lehre ihn den Gebrauch, wenn der Platz schon vorbereitet ist.

Und er ist es hier nicht dadurch, daß der, dem wir die Erklärung geben, schon Regeln weiß, sondern dadurch, daß er in anderm Sinne schon ein Spiel beherrscht.

Betrachte noch diesen Fall: Ich erkläre jemandem das Schachspiel; und fange damit an, indem ich auf eine Figur zeige und sage: »Das ist der König. Er kann so und so ziehen, etc. etc.«.— In diesem Fall werden wir sagen: die Worte »Das ist der König« (oder »Das heißt ›König‹«) sind nur dann eine Worterklärung, wenn der Lernende schon ›weiß, was eine Spielfigur ist‹. Wenn er also etwa schon andere Spiele gespielt hat, oder dem Spielen Anderer ›mit Verständnis‹ zugesehen hat—*und dergleichen.* Auch nur dann wird er beim Lernen des Spiels relevant fragen können: »Wie heißt das?«—nämlich, diese Spielfigur.

Wir können sagen: Nach der Benennung fragt nur der sinnvoll, der schon etwas mit ihr anzufangen weiß.

Wir können uns ja auch denken, daß der Gefragte antwortet: »Bestimm die Benennung selber«—und nun müßte, der gefragt hat, für alles selber aufkommen.

32. Wer in ein fremdes Land kommt, wird manchmal die Sprache der Einheimischen durch hinweisende Erklärungen lernen, die sie ihm geben; und er wird die Deutung dieser Erklärungen oft *raten* müssen und manchmal richtig, manchmal falsch raten.

Und nun können wir, glaube ich, sagen: Augustinus beschreibe das Lernen der menschlichen Sprache so, als käme das Kind in ein fremdes Land und verstehe die Sprache des Landes nicht; das heißt: so als habe es bereits eine Sprache, nur nicht diese. Oder auch: als könne das Kind schon *denken,* nur noch nicht sprechen. Und »denken« hieße hier etwas, wie: zu sich selber reden.

Russel, B.

1905

ON DENOTING

BY a 'denoting phrase' I mean a phrase such as any one of the following: a man, some man, any man, every man, all men, the present King of England, the present King of France, the centre of mass of the solar system at the first instant of the twentieth century, the revolution of the earth round the sun, the revolution of the sun round the earth. Thus a phrase is denoting solely in virtue of its *form*. We may distinguish three cases: (1) A phrase may be denoting, and yet not denote anything; e.g., 'the present King of France'. (2) A phrase may denote one definite object; e.g., 'the present King of England' denotes a certain man. (3) A phrase may denote ambiguously; e.g., 'a man' denotes not many men, but an ambiguous man. The interpretation of such phrases is a matter of considerable difficulty; indeed, it is very hard to frame any theory not susceptible of formal refutation. All the difficulties with which I am acquainted are met, so far as I can discover, by the theory which I am about to explain.

The subject of denoting is of very great importance, not only in logic and mathematics, but also in theory of knowledge. For example, we know that the centre of mass of the solar system at a definite instant is some definite point, and we can affirm a number of propositions about it; but we have no immediate *acquaintance* with this point, which is only known to us by description. The distinction between *acquaintance* and *knowledge about* is the distinction between the things we have presentations of, and the things we only reach by means of denoting phrases. It often happens that we know that a certain phrase denotes unambiguously, although we have no acquaintance with what it denotes; this occurs in the above case of the centre of mass. In perception we have acquaintance with the objects of perception, and in thought we have acquaintance with objects of a more abstract logical character;

but we do not necessarily have acquaintance with the objects denoted by phrases composed of words with whose meanings we are acquainted. To take a very important instance: there seems no reason to believe that we are ever acquainted with other people's minds, seeing that these are not directly perceived; hence what we know about them is obtained through denoting. All thinking has to start from acquaintance; but it succeeds in thinking *about* many things with which we have no acquaintance.

The course of my argument will be as follows. I shall begin by stating the theory I intend to advocate;* I shall then discuss the theories of Frege and Meinong, showing why neither of them satisfies me; then I shall give the grounds in favour of my theory; and finally I shall briefly indicate the philosophical consequences of my theory.

My theory, briefly, is as follows. I take the notion of the *variable* as fundamental; I use '$C(x)$' to mean a proposition† in which x is a constituent, where x, the variable, is essentially and wholly undetermined. Then we can consider the two notions '$C(x)$ is always true' and '$C(x)$ is sometimes true'‡. Then *everything* and *nothing* and *something* (which are the most primitive of denoting phrases) are to be interpreted as follows:

C (everything) means '$C(x)$ is always true';
C (nothing) means ' "$C(x)$ is false" is always true';
C (something) means 'It is false that "$C(x)$ is false" is always true'.§

Here the notion '$C(x)$ is always true' is taken as ultimate and indefinable, and the others are defined by means of it. *Everything*, *nothing*, and *something* are not assumed to have any meaning in isolation, but a meaning is assigned to *every* proposition in which they occur. This is the principle of the theory of denoting I wish

* I have discussed this subject in *Principles of Mathematics*, Chap. V, and § 476. The theory there advocated is very nearly the same as Frege's, and is quite different from the theory to be advocated in what follows.

† More exactly, a propositional function.

‡ The second of these can be defined by means of the first, if we take it to mean, 'It is not true that "$C(x)$ is false" is always true'.

§ I shall sometimes use, instead of this complicated phrase, the phrase '$C(x)$ is not always false', or '$C(x)$ is sometimes true', supposed *defined* to mean the same as the complicated phrase.

to advocate: that denoting phrases never have any meaning in themselves, but that every proposition in whose verbal expression they occur has a meaning. The difficulties concerning denoting are, I believe, all the result of a wrong analysis of propositions whose verbal expressions contain denoting phrases. The proper analysis, if I am not mistaken, may be further set forth as follows.

Suppose now we wish to interpret the proposition, 'I met a man'. It this is true, I met some definite man; but that is not what I affirm. What I affirm is, according to the theory I advocate:

' "I met x, and x is human" is not always false'.

Generally, defining the class of men as the class of objects having the predicate *human*, we say that:

'C (a man)' means ' "$C(x)$ and x is human" is not always false'. This leaves 'a man', by itself, wholly destitute of meaning, but gives a meaning to every proposition in whose verbal expression 'a man' occurs.

Consider next the proposition 'all men are mortal'. This proposition* is really hypothetical and states that *if* anything is a man, it is mortal. That is, it states that if x is a man, x is mortal, whatever x may be. Hence, substituting 'x is human' for 'x is a man', we find:

'All men are mortal' means ' "If x is human, x is mortal" is always true'.

This is what is expressed in symbolic logic by saying that 'all men are mortal' means ' "x is human" implies "x is mortal" for all values of x'. More generally, we say:

'C (all men)' means ' "If x is human, then $C(x)$ is true" is always true'.

Similarly

'C (no men)' means ' "If x is human, then $C(x)$ is false" is always true'.

'C (some men)' will mean the same as 'C (a man)',† and

* As has been ably argued in Mr. Bradley's *Logic*, Book I, Chap. II.

† Psychologically 'C (a man)' has a suggestion of *only one*, and 'C (some men)' has a suggestion of *more than one*; but we may neglect these suggestions in a preliminary sketch.

'C (a man)' means 'It is false that "$C(x)$ and x is human" is always false'.

'C (every man)' will mean the same as 'C (all men)'.

It remains to interpret phrases containing *the*. These are by far the most interesting and difficult of denoting phrases. Take as an instance 'the father of Charles II was executed'. This asserts that there was an x who was the father of Charles II and was executed. Now *the*, when it is strictly used, involves uniqueness; we do, it is true, speak of '*the* son of So-and-so' even when So-and-so has several sons, but it would be more correct to say '*a* son of So-and-so'. Thus for our purposes we take *the* as involving uniqueness. Thus when we say 'x was *the* father of Charles II' we not only assert that x had a certain relation to Charles II, but also that nothing else had this relation. The relation in question, without the assumption of uniqueness, and without any denoting phrases, is expressed by 'x begat Charles II'. To get an equivalent of 'x was the father of Charles II', we must add, 'If y is other than x, y did not beget Charles II', or, what is equivalent, 'If y begat Charles II, y is identical with x'. Hence 'x is the father of Charles II' becomes: 'x begat Charles II; and "if y begat Charles II, y is identical with x" is always true of y'.

Thus 'the father of Charles II was executed' becomes: 'It is not always false of x that x begat Charles II and that x was executed and that "if y begat Charles II, y is identical with x" is always true of y'.

This may seem a somewhat incredible interpretation; but I am not at present giving reasons, I am merely *stating* the theory.

To interpret 'C (the father of Charles II)', where C stands for any statement about him, we have only to substitute $C(x)$ for 'x was executed' in the above. Observe that, according to the above interpretation, whatever statement C may be, 'C (the father of Charles II)' implies:

'It is not always false of x that "if y begat Charles II, y is identical with x" is always true of y',

which is what is expressed in common language by 'Charles II had one father and no more'. Consequently if this condition fails, *every* proposition of the form 'C (the father of Charles II)' is false. Thus e.g. every proposition of the form 'C (the present King of France)' is false. This is a great advantage in the present theory. I shall show later that it is not contrary to the law of contradiction, as might be at first supposed.

The above gives a reduction of all propositions in which denoting phrases occur to forms in which no such phrases occur. Why it is imperative to effect such a reduction, the subsequent discussion will endeavour to show.

The evidence for the above theory is derived from the difficulties which seem unavoidable if we regard denoting phrases as standing for genuine constituents of the propositions in whose verbal expressions they occur. Of the possible theories which admit such constituents the simplest is that of Meinong.* This theory regards any grammatically correct denoting phrase as standing for an *object*. Thus 'the present King of France', 'the round square', etc., are supposed to be genuine objects. It is admitted that such objects do not *subsist*, but nevertheless they are supposed to be objects. This is in itself a difficult view; but the chief objection is that such objects, admittedly, are apt to infringe the law of contradiction. It is contended, for example, that the existent present King of France exists, and also does not exist; that the round square is round, and also not round, etc. But this is intolerable; and if any theory can be found to avoid this result, it is surely to be preferred.

The above breach of the law of contradiction is avoided by Frege's theory. He distinguishes, in a denoting phrase, two elements, which we may call the *meaning* and the *denotation*.† Thus 'the centre of mass of the solar system at the beginning of the twentieth century' is highly complex in *meaning*, but its *denotation* is a certain point, which is simple. The solar system, the twentieth century, etc., are constituents of the *meaning*; but the *denotation*

* See *Untersuchungen zur Gegenstandstheorie und Psychologie* (Leipzig, 1904) the first three articles (by Meinong, Ameseder and Mally respectively).

† See his 'Ueber Sinn und Bedeutung', *Zeitschrift für Phil. und Phil. Kritik*, Vol. 100.

has no constituents at all.* One advantage of this distinction is that it shows why it is often worth while to assert identity. If we say 'Scott is the author of *Waverley*', we assert an identity of denotation with a difference of meaning. I shall, however, not repeat the grounds in favour of this theory, as I have urged its claims elsewhere (loc. cit.), and am now concerned to dispute those claims.

One of the first difficulties that confront us, when we adopt the view that denoting phrases *express* a meaning and *denote* a denotation,† concerns the cases in which the denotation appears to be absent. If we say 'the King of England is bald', that is, it would seem, not a statement about the complex *meaning* 'the King of England', but about the actual man denoted by the meaning. But now consider 'the King of France is bald'. By parity of form, this also ought to be about the denotation of the phrase ' the King of France'. But this phrase, though it has a *meaning* provided 'the King of England' has a meaning, has certainly no denotation, at least in any obvious sense. Hence one would suppose that 'the King of France is bald' ought to be nonsense; but it is not nonsense, since it is plainly false. Or again consider such a proposition as the following: 'If u is a class which has only one member, then that one member is a member of u', or, as we may state it, 'If u is a unit class, *the* u is a u'. This proposition ought to be *always* true, since the conclusion is true whenever the hypothesis is true. But 'the u' is a denoting phrase, and it is the denotation, not the meaning, that is said to be a u. Now if u is *not* a unit class, 'the u' seems to denote nothing; hence our proposition would seem to become nonsense as soon as u is not a unit class.

Now it is plain that such propositions do *not* become non-

* Frege distinguishes the two elements of meaning and denotation everywhere, and not only in complex denoting phrases. Thus it is the *meanings* of the constituents of a denoting complex that enter into its *meaning*, not their *denotation*. In the proposition 'Mont Blanc is over 1,000 metres high', it is, according to him, the *meaning* of 'Mont Blanc', not the actual mountain, that is a constituent of the *meaning* of the proposition.

† In this theory, we shall say that the denoting phrase *expresses* a meaning; and we shall say both of the phrase and of the meaning that they *denote* a denotation. In the other theory, which I advocate, there is no *meaning*, and only sometimes a *denotation*.

sense merely because their hypotheses are false. The King in *The Tempest* might say, 'If Ferdinand is not drowned, Ferdinand is my only son'. Now 'my only son' is a denoting phrase, which, on the face of it, has a denotation when, and only when, I have exactly one son. But the above statement would nevertheless have remained true if Ferdinand had been in fact drowned. Thus we must either provide a denotation in cases in which it is at first sight absent, or we must abandon the view that the denotation is what is concerned in propositions which contain denoting phrases. The latter is the course that I advocate. The former course may be taken, as by Meinong, by admitting objects which do not subsist, and denying that they obey the law of contradiction; this, however, is to be avoided if possible. Another way of taking the same course (so far as our present alternative is concerned) is adopted by Frege, who provides by definition some purely conventional denotation for the cases in which otherwise there would be none. Thus 'the King of France', is to denote the null-class; 'the only son of Mr. So-and-so' (who has a fine family of ten), is to denote the class of all his sons; and so on. But this procedure, though it may not lead to actual logical error, is plainly artificial, and does not give an exact analysis of the matter. Thus if we allow that denoting phrases, in general, have the two sides of meaning and denotation, the cases where there seems to be no denotation cause difficulties both on the assumption that there really is a denotation and on the assumption that there really is none.

A logical theory may be tested by its capacity for dealing with puzzles, and it is a wholesome plan, in thinking about logic, to stock the mind with as many puzzles as possible, since these serve much the same purpose as is served by experiments in physical science. I shall therefore state three puzzles which a theory as to denoting ought to be able to solve; and I shall show later that my theory solves them.

(1) If a is identical with b, whatever is true of the one is true of the other, and either may be substituted for the other in any proposition without altering the truth or falsehood of that proposition. Now George IV wished to know whether Scott was the author of *Waverley*; and in fact Scott *was* the author of *Waverley*. Hence we may substitute *Scott* for *the author of 'Waverley'*, and thereby prove that George IV wished to know whether Scott was Scott.

Yet an interest in the law of identity can hardly be attributed to the first gentleman of Europe.

(2) By the law of excluded middle, either '*A* is *B*' or '*A* is not *B*' must be true. Hence either 'the present King of France is bald' or 'the present King of France is not bald' must be true. Yet if we enumerated the things that are bald, and then the things that are not bald, we should not find the present King of France in either list. Hegelians, who love a synthesis, will probably conclude that he wears a wig.

(3) Consider the proposition '*A* differs from *B*'. If this is true, there is a difference between *A* and *B*, which fact may be expressed in the form 'the difference between *A* and *B* subsists'. But if it is false that *A* differs from *B*, then there is no difference between *A* and *B*, which fact may be expressed in the form 'the difference between *A* and *B* does not subsist'. But how can a non-entity be the subject of a proposition? 'I think, therefore I am' is no more evident than 'I am the subject of a proposition, therefore I am', provided 'I am' is taken to assert subsistence or being,* not existence. Hence, it would appear, it must always be self-contradictory to deny the being of anything; but we have seen, in connexion with Meinong, that to admit being also sometimes leads to contradictions. Thus if *A* and *B* do not differ, to suppose either that there is, or that there is not, such an object as 'the difference between *A* and *B*' seems equally impossible.

The relation of the meaning to the denotation involves certain rather curious difficulties, which seem in themselves sufficient to prove that the theory which leads to such difficulties must be wrong.

When we wish to speak about the *meaning* of a denoting phrase, as opposed to its *denotation*, the natural mode of doing so is by inverted commas. Thus we say:

The centre of mass of the solar system is a point, not a denoting complex;

'The centre of mass of the solar system' is a denoting complex, not a point.

Or again,

The first line of Gray's Elegy states a proposition.

* I use these as synonyms.

'The first line of Gray's Elegy' does not state a proposition. Thus taking any denoting phrase, say *C*, we wish to consider the relation between *C* and '*C*', where the difference of the two is of the kind exemplified in the above two instances.

We say, to begin with, that when *C* occurs it is the *denotation* that we are speaking about; but when '*C*' occurs, it is the *meaning*. Now the relation of meaning and denotation is not merely linguistic through the phrase: there must be a logical relation involved, which we express by saying that the meaning denotes the denotation. But the difficulty which confronts us is that we cannot succeed in *both* preserving the connexion of meaning and denotation *and* preventing them from being one and the same; also that the meaning cannot be got at except by means of denoting phrases. This happens as follows.

The one phrase *C* was to have both meaning and denotation. But if we speak of 'the meaning of *C*', that gives us the meaning (if any) of the denotation. 'The meaning of the first line of Gray's Elegy' is the same as 'The meaning of "The curfew tolls the knell of parting day",' and is not the same as 'The meaning of "the first line of Gray's Elegy".' Thus in order to get the meaning we want, we must speak not of 'the meaning of *C*', but of 'the meaning of "*C*",' which is the same as '*C*' by itself. Similarly 'the denotation of *C*' does not mean the denotation we want, but means something which, if it denotes at all, denotes what is denoted by the denotation we want. For example, let '*C*' be 'the denoting complex occurring in the second of the above instances'. Then

C = 'the first line of Gray's Elegy', and

the denotation of C = The curfew tolls the knell of parting day. But what we *meant* to have as the denotation was 'the first line of Gray's Elegy'. Thus we have failed to get what we wanted.

The difficulty in speaking of the meaning of a denoting complex may be stated thus: The moment we put the complex in a proposition, the proposition is about the denotation; and if we make a proposition in which the subject is 'the meaning of *C*', then the subject is the meaning (if any) of the denotation, which was not intended. This leads us to say that, when we distinguish meaning and denotation, we must be dealing with the meaning: the meaning has denotation and is a complex, and there is not something

other than the meaning, which can be called the complex, and be said to *have* both meaning and denotation. The right phrase, on the view in question, is that some meanings have denotations.

But this only makes our difficulty in speaking of meanings more evident. For suppose C is our complex; then we are to say that C *is* the meaning of the complex. Nevertheless, whenever C occurs without inverted commas, what is said is not true of the meaning, but only of the denotation, as when we say: The centre of mass of the solar system is a point. Thus to speak of C itself, i.e., to make a proposition about the meaning, our subject must not be C, but something which denotes C. Thus 'C', which is what we use when we want to speak of the meaning, must be not the meaning, but something which denotes the meaning. And C must not be a constituent of this complex (as it is of 'the meaning of C'); for if C occurs in the complex, it will be its denotation, not its meaning, that will occur, and there is no backward road from denotations to meanings, because every object can be denoted by an infinite number of different denoting phrases.

Thus it would seem that 'C' and C are different entities, such that 'C' denotes C; but this cannot be an explanation, because the relation of 'C' to C remains wholly mysterious; and where are we to find the denoting complex 'C' which is to denote C? Moreover, when C occurs in a proposition, it is not *only* the denotation that occurs (as we shall see in the next paragraph); yet, on the view in question, C is only the denotation, the meaning being wholly relegated to 'C'. This is an inextricable tangle, and seems to prove that the whole distinction of meaning and denotation has been wrongly conceived.

That the meaning is relevant when a denoting phrase occurs in a proposition is formally proved by the puzzle about the author of *Waverley*. The proposition 'Scott was the author of *Waverley*' has a property not possessed by 'Scott was Scott', namely the property that George IV wished to know whether it was true. Thus the two are not identical propositions; hence the meaning of 'the author of *Waverley*' must be relevant as well as the denotation, if we adhere to the point of view to which this distinction belongs. Yet, as we have just seen, so long as we adhere to this point of view, we are compelled to hold that only the denotation

can be relevant. Thus the point of view in question must be abandoned.

It remains to show how all the puzzles we have been considering are solved by the theory explained at the beginning of this article.

According to the view which I advocate, a denoting phrase is essentially *part* of a sentence, and does not, like most single words, have any significance on its own account. If I say 'Scott was a man', that is a statement of the form 'x was a man', and it has 'Scott' for its subject. But if I say 'the author of *Waverley* was a man', that is not a statement of the form 'x was a man', and does not have 'the author of *Waverley*' for its subject. Abbreviating the statement made at the beginning of this article, we may put, in place of 'the author of *Waverley* was a man', the following: 'One and only one entity wrote *Waverley*, and that one was a man'. (This is not so strictly what is meant as what was said earlier; but it is easier to follow.) And speaking generally, suppose we wish to say that the author of *Waverley* had the property ϕ, what we wish to say is equivalent to 'One and only one entity wrote *Waverley*, and that one had the property ϕ'.

The explanation of *denotation* is now as follows. Every proposition in which 'the author of *Waverley*' occurs being explained as above, the proposition 'Scott was the author of *Waverley*' (i.e. 'Scott was identical with the author of *Waverley*') becomes 'One and only one entity wrote *Waverley*, and Scott was identical with that one'; or, reverting to the wholly explicit form: 'It is not always false of x that x wrote *Waverley*, that it is always true of y that if y wrote *Waverley* y is identical with x, and that Scott is identical with x'. Thus if 'C' is a denoting phrase, it may happen that there is one entity x (there cannot be more than one) for which the proposition 'x is identical with C' is true, this proposition being interpreted as above. We may then say that the entity x is the denotation of the phrase 'C'. Thus Scott is the denotation of 'the author of *Waverley*'. The 'C' in inverted commas will be merely the *phrase*, not anything that can be called the *meaning*. The phrase *per se* has no meaning, because in any proposition in which it occurs the proposition, fully expressed, does not contain the phrase, which has been broken up.

The puzzle about George IV's curiosity is now seen to have a very simple solution. The proposition 'Scott was the author of

Waverley', which was written out in its unabbreviated form in the preceding paragraph, does not contain any constituent 'the author of *Waverley*' for which we could substitute 'Scott'. This does not interfere with the truth of inferences resulting from making what is *verbally* the substitution of 'Scott' for 'the author of *Waverley*', so long as 'the author of *Waverley*' has what I call a *primary* occurrence in the proposition considered. The difference of primary and secondary occurrences of denoting phrases is as follows:

When we say: 'George IV wished to know whether so-and-so', or when we say 'So-and-so is surprising' or 'So-and-so is true', etc., the 'so-and-so' must be a proposition. Suppose now that 'so-and-so' contains a denoting phrase. We may either eliminate this denoting phrase from the subordinate proposition 'so-and-so', or from the whole proposition in which 'so-and-so' is a mere constituent. Different propositions result according to which we do. I have heard of a touchy owner of a yacht to whom a guest, on first seeing it, remarked, 'I thought your yacht was larger than it is'; and the owner replied, 'No, my yacht is not larger than it is'. What the guest meant was, 'The size that I thought your yacht was is greater than the size your yacht is'; the meaning attributed to him is, 'I thought the size of your yacht was greater than the size of your yacht'. To return to George IV and *Waverley*, when we say, 'George IV wished to know whether Scott was the author of *Waverley*', we normally mean 'George IV wished to know whether one and only one man wrote *Waverley* and Scott was that man'; but we *may* also mean: 'One and only one man wrote *Waverley*, and George IV wished to know whether Scott was that man'. In the latter, 'the author of *Waverley*' has a *primary* occurrence; in the former, a *secondary*. The latter might be expressed by 'George IV wished to know, concerning the man who in fact wrote *Waverley*, whether he was Scott'. This would be true, for example, if George IV had seen Scott at a distance, and had asked 'Is that Scott?'. A *secondary* occurrence of a denoting phrase may be defined as one in which the phrase occurs in a proposition p which is a mere constituent of the proposition we are considering, and the substitution for the denoting phrase is to be effected in p, not in the whole proposition concerned. The ambiguity as between primary and secondary occurrences is hard to avoid in language;

but it does no harm if we are on our guard against it. In symbolic logic it is of course easily avoided.

The distinction of primary and secondary occurrences also enables us to deal with the question whether the present King of France is bald or not bald, and generally with the logical status of denoting phrases that denote nothing. If 'C' is a denoting phrase, say 'the term having the property F', then

'C has the property ϕ' means 'one and only one term has the property F, and that one has the property ϕ'.*

If now the property F belongs to no terms, or to several, it follows that 'C has the property ϕ' is false for *all* values of ϕ. Thus 'the present King of France is bald' is certainly false; and 'the present King of France is not bald' is false if it means

'There is an entity which is now King of France and is not bald', but is true if it means

'It is false that there is an entity which is now King of France and is bald'.

That is, 'the King of France is not bald' is false if the occurrence of 'the King of France' is *primary*, and true if it is *secondary*. Thus all propositions in which 'the King of France' has a primary occurrence are false; the denials of such propositions are true, but in them 'the King of France' has a secondary occurrence. Thus we escape the conclusion that the King of France has a wig.

We can now see also how to deny that there is such an object as the difference between A and B in the case when A and B do not differ. If A and B do differ, there is one and only one entity x such that 'x is the difference between A and B' is a true proposition; if A and B do not differ, there is no such entity x. Thus according to the meaning of denotation lately explained, 'the difference between A and B' has a denotation when A and B differ, but not otherwise. This difference applies to true and false propositions generally. If '$a\,R\,b$' stands for 'a has the relation R to b', then when $a\,R\,b$ is true, there is such an entity as the relation R between a and b; when $a\,R\,b$ is false, there is no such entity. Thus out of any proposition we can make a denoting phrase, which denotes an entity if the proposition is true, but does not denote an entity

* This is the abbreviated, not the stricter, interpretation.

if the proposition is false. E.g., it is true (at least we will suppose so) that the earth revolves round the sun, and false that the sun revolves round the earth; hence 'the revolution of the earth round the sun' denotes an entity, while 'the revolution of the sun round the earth' does not denote an entity.*

The whole realm of non-entities, such as 'the round square', 'the even prime other than 2', 'Apollo', 'Hamlet', etc., can now be satisfactorily dealt with. All these are denoting phrases which do not denote anything. A proposition about Apollo means what we get by substituting what the classical dictionary tells us is meant by Apollo, say 'the sun-god'. All propositions in which Apollo occurs are to be interpreted by the above rules for denoting phrases. If 'Apollo' has a primary occurrence, the proposition containing the occurrence is false; if the occurrence is secondary, the proposition may be true. So again 'the round square is round' means 'there is one and only one entity x which is round and square, and that entity is round', which is a false proposition, not, as Meinong maintains, a true one. 'The most perfect Being has all perfections; existence is a perfection; therefore the most perfect Being exists' becomes:

'There is one and only one entity x which is most perfect; that one has all perfections; existence is a perfection; therefore that one exists'. As a proof, this fails for want of a proof of the premiss 'there is one and only one entity x which is most perfect'.†

Mr. MacColl (*Mind*, N.S., No. 54, and again No. 55, page 401) regards individuals as of two sorts, real and unreal; hence he defines the null-class as the class consisting of all unreal individuals. This assumes that such phrases as 'the present King of France', which do not denote a real individual, do, nevertheless, denote an individual, but an unreal one. This is essentially Meinong's theory, which we have seen reason to reject because it conflicts with the law of contradiction. With our theory of denoting, we are able to

* The propositions from which such entities are derived are not identical either with these entities or with the propositions that these entities have being.

† The argument can be made to prove validly that all members of the class of most perfect Beings exist; it can also be proved formally that this class cannot have *more* than one member; but, taking the definition of perfection as possession of all positive predicates, it can be proved almost equally formally that the class does not have even one member.

hold that there are no unreal individuals; so that the null-class is the class containing no members, not the class containing as members all unreal individuals.

It is important to observe the effect of our theory on the interpretation of definitions which proceed by means of denoting phrases. Most mathematical definitions are of this sort; for example '$m-n$ means the number which, added to n, gives m'. Thus $m-n$ is defined as meaning the same as a certain denoting phrase; but we agreed that denoting phrases have no meaning in isolation. Thus what the definition really ought to be is: 'Any proposition containing $m-n$ is to mean the proposition which results from substituting for "$m-n$" "the number which, added to n, gives m".' The resulting proposition is interpreted according to the rules already given for interpreting propositions whose verbal expression contains a denoting phrase. In the case where m and n are such that there is one and only one number x which, added to n, gives m, there is a number x which can be substituted for $m-n$ in any proposition containing $m-n$ without altering the truth or falsehood of the proposition. But in other cases, all propositions in which '$m-n$' has a primary occurrence are false.

The usefulness of *identity* is explained by the above theory. No one outside a logic-book ever wishes to say 'x is x', and yet assertions of identity are often made in such forms as 'Scott was the author of *Waverley*' or 'thou art the man'. The meaning of such propositions cannot be stated without the notion of identity, although they are not simply statements that Scott is identical with another term, the author of *Waverley*, or that thou art identical with another term, the man. The shortest statement of 'Scott is the author of *Waverley*' seems to be 'Scott wrote *Waverley*; and it is always true of y that if y wrote *Waverley*, y is identical with Scott'. It is in this way that identity enters into 'Scott is the author of *Waverley*'; and it is owing to such uses that identity is worth affirming.

One interesting result of the above theory of denoting is this: when there is anything with which we do not have immediate acquaintance, but only definition by denoting phrases, then the propositions in which this thing is introduced by means of a denoting phrase do not really contain this thing as a constituent, but contain instead the constituents expressed by the several words

of the denoting phrase. Thus in every proposition that we can apprehend (i.e. not only in those whose truth or falsehood we can judge of, but in all that we can think about), all the constituents are really entities with which we have immediate acquaintance. Now such things as matter (in the sense in which matter occurs in physics) and the minds of other people are known to us only by denoting phrases, i.e. we are not *acquainted* with them, but we know them as what has such and such properties. Hence, although we can form propositional functions $C(x)$ which must hold of such and such a material particle, or of So-and-so's mind, yet we are not acquainted with the propositions which affirm these things that we know must be true, because we cannot apprehend the actual entities concerned. What we know is 'So-and-so has a mind which has such and such properties' but we do not know 'A has such and such properties', where A *is* the mind in question. In such a case, we know the properties of a thing without having acquaintance with the thing itself, and without, consequently, knowing any single proposition of which the thing itself is a constituent.

Of the many other consequences of the view I have been advocating, I will say nothing. I will only beg the reader not to make up his mind against the view—as he might be tempted to do, on account of its apparently excessive complication—until he has attempted to construct a theory of his own on the subject of denotation. This attempt, I believe, will convince him that, whatever the true theory may be, it cannot have such a simplicity as one might have expected beforehand.

KRIPKE, SAUL

NAMING AND NECESSITY

LECTURES I + II

LECTURE I: JANUARY 20, 1970[1]

I hope that some people see some connection between the two topics in the title. If not, anyway, such connections will be developed in the course of these talks. Furthermore, because of the use of tools involving reference and necessity in analytic philosophy today, our views on these topics really have wide-

[1] In January of 1970, I gave three talks at Princeton University transcribed here. As the style of the transcript makes clear, I gave the talks without a written text, and, in fact, without notes. The present text is lightly edited from the *verbatim* transcript; an occasional passage has been added to expand the thought, an occasional sentence has been rewritten, but no attempt has been made to change the informal style of the original. Many of the footnotes have been added to the original, but a few were originally spoken asides in the talks themselves.

I hope the reader will bear these facts in mind as he reads the text. Imagining it spoken, with proper pauses and emphases, may occasionally facilitate comprehension. I have agreed to publish the talks in this form with some reservations. The time allotted, and the informal style, necessitated a certain amount of compression of the argument, inability to treat certain objections, and the like. Especially in the concluding sections on scientific identities and the mind-body problem thoroughness had to be sacrificed. Some topics essential to a full presentation of the viewpoint argued here, especially that of existence statements and empty names, had to be omitted altogether. Further, the informality of the presentation may well have engendered a sacrifice of clarity at certain points. All these defects were accepted in the interest of early publication. I hope that perhaps I will have the chance to do a more thorough job later. To repeat, I hope the reader will bear in mind that he is largely reading informal lectures, not only when he encounters repetitions or infelicities, but also when he encounters irreverence or corn.

ranging implications for other problems in philosophy that traditionally might be thought far-removed, like arguments over the mind-body problem or the so-called 'identity thesis'. Materialism, in this form, often now gets involved in very intricate ways in questions about what is necessary or contingent in identity of properties—questions like that. So, it is really very important to philosophers who may want to work in many domains to get clear about these concepts. Maybe I will say something about the mind-body problem in the course of these talks. I want to talk also at some point (I don't know if I can get it in) about substances and natural kinds.

The way I approach these matters will be, in some ways, quite different from what people are thinking today (though it also has some points of contact with what some people have been thinking and writing today, and if I leave people out in informal talks like this, I hope that I will be forgiven).[2] Some of the views that I have are views which may at first glance strike some as obviously wrong. My favorite example is this (which I probably won't defend in the lectures—for one thing it doesn't ever convince anyone): It is a common claim in contemporary philosophy that there are certain predicates which, though they are in fact empty—have null extension—have it

[2] Given a chance to add a footnote, I shall mention that Rogers Albritton, Charles Chastain, Keith Donnellan, and Michael Slote (in addition to philosophers mentioned in the text, especially Hilary Putnam), have independently expressed views with points of contact with various aspects of what I say here. Albritton called the problems of necessity and *a prioricity* in natural kinds to my attention, by raising the question whether we could discover that lemons were not fruits. (I am not sure he would accept all my conclusions.) I also recall the influence of early conversations with Albritton and with Peter Geach on the essentiality of origins. The apology in the text still stands; I am aware that the list in this footnote is far from comprehensive. I make no attempt to enumerate those friends and students whose stimulating conversations have helped me. Thomas Nagel and Gilbert Harman deserve special thanks for their help in editing the transcript.

as a matter of contingent fact and not as a matter of any sort of necessity. Well, *that* I don't dispute; but an example which is usually given is the example of *unicorn*. So it is said that though we have all found out that there are no unicorns, of course there *might* have been unicorns. Under certain circumstances there *would* have been unicorns. And this is an example of something I think is not the case. Perhaps according to me the truth should not be put in terms of saying that it is necessary that there should be no unicorns, but just that we can't say under what circumstances there would have been unicorns. Further, I think that even if archeologists or geologists were to discover tomorrow some fossils conclusively showing the existence of animals in the past satisfying everything we know about unicorns from the myth of the unicorn, that would not show that there were unicorns. Now I don't know if I'm going to have a chance to defend this particular view, but it's an example of a surprising one. (I actually gave a seminar in this institution where I talked about this view for a couple of sessions.) So, some of my opinions are somewhat surprising; but let us start out with some area that is perhaps not as surprising and introduce the methodology and problems of these talks.

The first topic in the pair of topics is naming. By a name here I will mean a proper name, i.e., the name of a person, a city, a country, etc. It is well known that modern logicians also are very interested in definite descriptions: phrases of the form 'the x such that φx', such as 'the man who corrupted Hadleyburg'. Now, if one and only one man ever corrupted Hadleyburg, then that man is the referent, in the logician's sense, of that description. We will use the term 'name' so that it does *not* include definite descriptions of that sort, but only those things which in ordinary language would be called 'proper names'. If we want a common term to cover names and descriptions, we may use the term 'designator'.

It is a point, made by Donnellan,[3] that under certain circumstances a particular speaker may use a definite description to refer, not to the proper referent, in the sense that I've just defined it, of that description, but to something else which he wants to single out and which he thinks is the proper referent of the description, but which in fact isn't. So you may say, 'The man over there with the champagne in his glass is happy', though he actually only has water in his glass. Now, even though there is no champagne in his glass, and there may be another man in the room who does have champagne in his glass, the speaker *intended* to refer, or maybe, in some sense of 'refer', *did* refer, to the man he thought had the champagne in his glass. Nevertheless, I'm just going to use the term 'referent

[3] Keith Donnellan, 'Reference and Definite Descriptions', *Philosophical Review* 75 (1966), pp. 281–304. See also Leonard Linsky, 'Reference and Referents', in *Philosophy and Ordinary Language* (ed. Caton), University of Illinois Press, Urbana, 1963. Donnellan's distinction seems applicable to names as well as to descriptions. Two men glimpse someone at a distance and think they recognize him as Jones. 'What is Jones doing?' 'Raking the leaves'. If the distant leaf-raker is actually Smith, then in some sense they are *referring* to Smith, even though they both use 'Jones' *as a name of* Jones. In the text, I speak of the 'referent' of a name to mean the thing named by the name—e.g., Jones, not Smith—even though a speaker may sometimes properly be said to use the name to refer to someone else. Perhaps it would have been less misleading to use a technical term, such as 'denote' rather than 'refer'. My use of 'refer' is such as to satisfy the schema, 'The referent of "X" is X', where 'X' is replaceable by any name or description. I am tentatively inclined to believe, in opposition to Donnellan, that his remarks about reference have little to do with semantics or truth-conditions, though they may be relevant to a theory of speech-acts. Space limitations do not permit me to explain what I mean by this, much less defend the view, except for a brief remark: Call the referent of a name or description in my sense the 'semantic referent'; for a name, this is the thing named, for a description, the thing uniquely satisfying the description.

Then the speaker may *refer* to something other than the semantic referent if he has appropriate false beliefs. I think this is what happens in the naming (Smith-Jones) cases and also in the Donnellan 'champagne' case; the one requires no theory that names are ambiguous, and the other requires no modification of Russell's theory of descriptions.

of the description' to mean the object uniquely satisfying the conditions in the definite description. This is the sense in which it's been used in the logical tradition. So, if you have a description of the form 'the x such that φx', and there is exactly one x such that φx, that is the referent of the description.

Now, what is the relation between names and descriptions? There is a well known doctrine of John Stuart Mill, in his book *A System of Logic*, that names have denotation but not connotation. To use one of his examples, when we use the name 'Dartmouth' to describe a certain locality in England, it may be so called because it lies at the mouth of the Dart. But even, he says, had the Dart (that's a river) changed its course so that Dartmouth no longer lay at the mouth of the Dart, we could still with propriety call this place 'Dartmouth', even though the name may suggest that it lies at the mouth of the Dart. Changing Mill's terminology, perhaps we should say that a name such as 'Dartmouth' *does* have a 'connotation' to some people, namely, it *does* connote (not to me—I never thought of this) that any place called 'Dartmouth' lies at the mouth of the Dart. But then in some way it doesn't have a 'sense'. At least, it is not part of the *meaning* of the name 'Dartmouth' that the town so named lies at the mouth of the Dart. Someone who said that Dartmouth did not lie at the Dart's mouth would not contradict himself.

It should not be thought that every phrase of the form 'the x such that Fx' is always used in English as a description rather than a name. I guess everyone has heard about The Holy Roman Empire, which was neither holy, Roman nor an empire. Today we have The United Nations. Here it would seem that since these things can be so-called even though they are not Holy Roman United Nations, these phrases should be regarded not as definite descriptions, but as names. In the case of some terms, people might have doubts as to whether they're names or descriptions; like 'God'—does it describe God as the

unique divine being or is it a name of God? But such cases needn't necessarily bother us.

Now here I am making a distinction which is certainly made in language. But the classical tradition of modern logic has gone very strongly against Mill's view. Frege and Russell both thought, and seemed to arrive at these conclusions independently of each other, that Mill was wrong in a very strong sense: really a proper name, properly used, simply was a definite description abbreviated or disguised. Frege specifically said that such a description gave the sense of the name.[4]

Now the reasons against Mill's view and in favor of the alternative view adopted by Frege and Russell are really very powerful; and it is hard to see—though one may be suspicious of this view because names don't seem to be disguised descriptions—how the Frege-Russell view, or some suitable variant, can fail to be the case.

Let me give an example of some of the arguments which seem conclusive in favor of the view of Frege and Russell. The basic problem for any view such as Mill's is how we can determine what the referent of a name, as used by a given

[4] Strictly speaking, of course, Russell says that the names don't abbreviate descriptions and don't have any sense; but then he also says that, just because the things that we call 'names' do abbreviate descriptions, they're not really names. So, since 'Walter Scott', according to Russell, does abbreviate a description, 'Walter Scott' is not a name; and the only names that really exist in ordinary language are, perhaps, demonstratives such as 'this' or 'that', used on a particular occasion to refer to an object with which the speaker is 'acquainted' in Russell's sense. Though we won't put things the way Russell does, we could describe Russell as saying that names, as they are ordinarily called, *do* have sense. They have sense in a strong way, namely, we should be able to give a definite description such that the referent of the name, by definition, is the object satisfying the description. Russell himself, since he eliminates descriptions from his primitive notation, seems to hold in 'On Denoting' that the notion of 'sense' is illusory. In reporting Russell's views, we thus deviate from him in two respects. First, we stipulate that 'names' shall be names as ordinarily conceived, not Russell's 'logically proper names'; second, we regard descriptions, and their abbreviations, as having sense.

speaker, is. According to the description view, the answer is clear. If 'Joe Doakes' is just short for 'the man who corrupted Hadleyburg', then whoever corrupted Hadleyburg uniquely is the referent of the name 'Joe Doakes'. However, if there is *not* such a descriptive content to the name, then how do people ever use names to refer to things at all? Well, they may be in a position to point to some things and thus determine the references of certain names ostensively. This was Russell's doctrine of acquaintance, which he thought the so-called genuine or proper names satisfied. But of course ordinary names refer to all sorts of people, like Walter Scott, to whom we can't possibly point. And our reference here seems to be determined by our knowledge of them. Whatever we know about them determines the referent of the name as the unique thing satisfying those properties. For example, if I use the name 'Napoleon', and someone asks, 'To whom are you referring?', I will answer something like, 'Napoleon was emperor of the French in the early part of the nineteenth century; he was eventually defeated at Waterloo', thus giving a uniquely identifying description to determine the referent of the name. Frege and Russell, then, appear to give the natural account of how reference is determined here; Mill appears to give none.

There are subsidiary arguments which, though they are based on more specialized problems, are also motivations for accepting the view. One is that sometimes we may discover that two names have the same referent, and express this by an identity statement. So, for example (I guess this is a hackneyed example), you see a star in the evening and it's called 'Hesperus'. (That's what we call it in the evening, is that right?—I hope it's not the other way around.) We see a star in the morning and call it 'Phosphorus'. Well, then, in fact we find that it's not a star, but is the planet Venus and that Hesperus and Phosphorus are in fact the same. So we express this by 'Hesperus is Phosphorus'. Here we're certainly not just saying of an object that it's identical with itself. This is something that we discovered. A very natural thing to say is that the real content [is that] the star which we saw in the evening is the star which we saw in the morning (or, more accurately, that the thing which we saw in the evening is the thing which we saw in the morning). This, then, gives the real meaning of the identity statement in question; and the analysis in terms of descriptions does this.

Also we may raise the question whether a name has any reference at all when we ask, e.g., whether Aristotle ever existed. It seems natural here to think that what is questioned is not whether this *thing* (man) existed. Once we've *got* the thing, we know that it existed. What really is queried is whether anything answers to the properties we associate with the name—in the case of Aristotle, whether any one Greek philosopher produced certain works, or at least a suitable number of them.

It would be nice to answer all of these arguments. I am not entirely able to see my way clear through every problem of this sort that can be raised. Furthermore, I'm pretty sure that I won't have time to discuss all these questions in these lectures. Nevertheless, I think it's pretty certain that the view of Frege and Russell is false.[5]

[5] When I speak of the Frege-Russell view and its variants, I include only those versions which give a substantive theory of the reference of names. In particular, Quine's proposal that in a 'canonical notation' a name such as 'Socrates' should be replaced by a description 'the Socratizer' (where 'Socratizes' is an invented predicate), and that the description should then be eliminated by Russell's method, was not intended as a theory of reference for names but as a proposed reform of language with certain advantages. The problems discussed here will all apply, *mutatis mutandis*, to the reformed language; in particular, the question, 'How is the reference of "Socrates" determined?' yields to the question, 'How is the extension of "Socratizes" determined?' Of course I do not suggest that Quine has ever claimed the contrary.

Many people have said that the theory of Frege and Russell is false, but, in my opinion, they have abandoned its letter while retaining its spirit, namely, they have used the notion of a cluster concept. Well, what is this? The obvious problem for Frege and Russell, the one which comes immediately to mind, is already mentioned by Frege himself. He said,

> In the case of genuinely proper names like 'Aristotle' opinions as regards their sense may diverge. As such may, e.g., be suggested: Plato's disciple and the teacher of Alexander the Great. Whoever accepts this sense will interpret the meaning of the statement 'Aristotle was born in Stagira', differently from one who interpreted the sense of 'Aristotle' as the Stagirite teacher of Alexander the Great. As long as the nominatum remains the same, these fluctuations in sense are tolerable. But they should be avoided in the system of a demonstrative science and should not appear in a perfect language.[6]

So, according to Frege, there is some sort of looseness or weakness in our language. Some people may give one sense to the name 'Aristotle', others may give another. But of course it is not only that; even a single speaker when asked 'What description are you willing to substitute for the name?' may be quite at a loss. In fact, he may know many things about him; but any particular thing that he knows he may feel clearly expresses a contingent property of the object. If 'Aristotle' meant *the man who taught Alexander the Great*, then saying 'Aristotle was a teacher of Alexander the Great' would be a mere tautology. But surely it isn't; it expresses the fact that Aristotle taught Alexander the Great, something we could discover to be false. So, *being the teacher of Alexander the Great* cannot be part of [the sense of] the name.

[6] Gottlob Frege, 'On Sense and Nominatum', translated by Herbert Feigl in *Readings in Philosophical Analysis* (ed. by Herbert Feigl and Wilfrid Sellars), Appleton Century Crofts, 1949, p. 86.

The most common way out of this difficulty is to say 'really it is not a weakness in ordinary language that we can't substitute a *particular* description for the name; that's all right. What we really associate with the name is a *family* of descriptions.' A good example of this is (if I can find it) in *Philosophical Investigations*, where the idea of family resemblances is introduced and with great power.

> Consider this example. If one says 'Moses did not exist', this may mean various things. It may mean: the Israelites did not have a *single* leader when they withdrew from Egypt—or: their leader was not called Moses—or: there cannot have been anyone who accomplished all that the Bible relates of Moses— . . . But when I make a statement about Moses,—am I always ready to substitute some *one* of those descriptions for 'Moses'? I shall perhaps say: by 'Moses' I understand the man who did what the Bible relates of Moses, or at any rate, a good deal of it. But how much? Have I decided how much must be proved false for me to give up my proposition as false? Has the name 'Moses' got a fixed and unequivocal use for me in all possible cases?[7]

According to this view, and a *locus classicus* of it is Searle's article on proper names,[8] the referent of a name is determined not by a single description but by some cluster or family. Whatever in some sense satisfies enough or most of the family is the referent of the name. I shall return to this view later. It may seem, as an analysis of ordinary language, quite a bit more plausible than that of Frege and Russell. It may seem to keep all the virtues and remove the defects of this theory.

Let me say (and this will introduce us to another new topic before I really consider this theory of naming) that there are two ways in which the cluster concept theory, or even the

[7] Ludwig Wittgenstein, *Philosophical Investigations*, translated by G. E. M. Anscombe, MacMillan, 1953, § 79.

[8] John R. Searle, 'Proper Names', *Mind* 67 (1958), 166–73.

theory which requires a single description, can be viewed. One way of regarding it says that the cluster or the single description actually gives the meaning of the name; and when someone says 'Walter Scott', he means *the man such that such and such and such and such.*

Now another view might be that even though the description in some sense doesn't give the *meaning* of the name, it is what *determines its reference* and although the phrase 'Walter Scott' isn't *synonymous* with 'the man such that such and such and such and such', or even maybe with the family (if something can be synonymous with a family), the family or the single description is what is used to determine to whom someone is referring when he says 'Walter Scott'. Of course, if when we hear his beliefs about Walter Scott we find that they are actually much more nearly true of Salvador Dali, then according to this theory the reference of this name is going to be Mr. Dali, not Scott. There are writers, I think, who explicitly deny that names have meaning at all even more strongly than I would but still use this picture of how the referent of the name gets determined. A good case in point is Paul Ziff, who says, very emphatically, that names don't have meaning at all, [that] they are not a part of language in some sense. But still, when he talks about how we determine what the reference of the name was, then he gives this picture. Unfortunately I don't have the passage in question with me, but this is what he says.[9]

[9] Ziff's most detailed statement of his version of the cluster-of-descriptions theory of the reference of names is in 'About God', reprinted in *Philosophical Turnings*, Cornell University Press, Ithaca, and Oxford University Press, London, 1966, pp. 94–96. A briefer statement is in his *Semantic Analysis*, Cornell University Press, Ithaca, 1960, pp. 102–05 (esp. pp. 103–04). The latter passage suggests that names of things with which we are acquainted should be treated somewhat differently (using ostension and baptism) from names of historical figures, where the reference is determined by (a cluster of) associated descriptions. On p. 93 of *Semantic Analysis* Ziff states that 'simple strong generalization(s) about proper names' are impossible; 'one can only say what is so for the most part . . .' Nevertheless Ziff clearly states that a cluster-of-descriptions theory is a reasonable such rough statement, at least for historical figures. For Ziff's view that proper names ordinarily are not words of the language and ordinarily do not have meaning, see pp. 85–89 and 93–94 of *Semantic Analysis*.

The difference between using this theory as a theory of meaning and using it as a theory of reference will come out a little more clearly later on. But some of the attractiveness of the theory is lost if it isn't supposed to give the meaning of the name; for some of the solutions of problems that I've just mentioned will not be right, or at least won't clearly be right, if the description doesn't give the meaning of the name. For example, if someone said 'Aristotle does not exist' *means* 'there is no man doing such and such', or in the example from Wittgenstein, 'Moses does not exist', *means* 'no man did such and such', that might depend (and in fact, I think, does depend) on taking the theory in question as a theory of the meaning of the name 'Moses', not just as a theory of its reference. Well, I don't know. Perhaps all that is immediate now is the other way around: if 'Moses' means the same as 'the man who did such and such' then to say that Moses did not exist is to say that the man who did such and such did not exist, that is, that no one person did such and such. If, on the other hand, 'Moses' is not synonymous with any description, then even if its reference is in some sense determined by a description, statements containing the name cannot in general be *analyzed* by replacing the name by a description, though they may be materially equivalent to statements containing a description. So the analysis of singular existence statements mentioned above will have to be given up, unless it is established by some special argument, independent of a general theory of the meaning of names; and the same applies to identity statements. In any case, I think it's false that 'Moses exists' means that at all.

So we won't have to see if such a special argument can be drawn up.[10]

Before I go any further into this problem, I want to talk about another distinction which will be important in the methodology of these talks. Philosophers have talked (and, of course, there has been considerable controversy in recent years over the meaningfulness of these notions) [about] various categories of truth, which are called '*a priori*', 'analytic', 'necessary'—and sometimes even 'certain' is thrown into this batch. The terms are often used as if *whether* there are things answering to these concepts is an interesting question, but we might as well regard them all as meaning the same thing. Now, everyone remembers Kant (a bit) as making a distinction between '*a priori*' and 'analytic'. So maybe this distinction is still made. In contemporary discussion very few people, if any, distinguish between the concepts of statements being *a priori* and their being necessary. At any rate I shall *not* use the terms '*a priori*' and 'necessary' interchangeably here.

Consider what the traditional characterizations of such terms as '*a priori*' and 'necessary' are. First the notion of a prioricity is a concept of epistemology. I guess the traditional characterization from Kant goes something like: *a priori* truths are those which can be known independently of any experience. This introduces another problem before we get off the ground, because there's another modality in the characterization of '*a priori*', namely, it is supposed to be something which *can* be known independently of any experience. That means that in some sense it's *possible* (whether we do or do not in fact know it independently of any experience) to know this independently of any experience. And possible for whom? For God? For the Martians?

[10] Those determinists who deny the importance of the individual in history may well argue that had Moses never existed, someone else would have arisen to achieve all that he did. Their claim cannot be refuted by appealing to a correct philosophical theory of the meaning of 'Moses exists'.

Or just for people with minds like ours? To make this all clear might [involve] a host of problems all of its own about what sort of possibility is in question here. It might be best therefore, instead of using the phrase '*a priori* truth', to the extent that one uses it at all, to stick to the question of whether a particular person or knower knows something *a priori* or believes it true on the basis of *a priori* evidence.

I won't go further too much into the problems that might arise with the notion of a prioricity here. I will say that some philosophers somehow change the modality in this characterization from *can* to *must*. They think that if something belongs to the realm of *a priori* knowledge, it couldn't possibly be known empirically. This is just a mistake. Something may belong in the realm of such statements that *can* be known *a priori* but still may be known by particular people on the basis of experience. To give a really common sense example: anyone who has worked with a computing machine knows that the computing machine may give an answer to whether such and such a number is prime. No one has calculated or proved that the number is prime; but the machine has given the answer: this number is prime. We, then, if we believe that the number is prime, believe it on the basis of our knowledge of the laws of physics, the construction of the machine, and so on. We therefore do not believe this on the basis of purely *a priori* evidence. We believe it (if anything is *a posteriori* at all) on the basis of *a posteriori* evidence. Nevertheless, maybe this could be known *a priori* by someone who made the requisite calculations. So '*can* be known *a priori*' doesn't mean '*must* be known *a priori*'.

The second concept which is in question is that of necessity. Sometimes this is used in an epistemological way and might then just mean *a priori*. And of course, sometimes it is used in a physical way when people distinguish between physical and logical necessity. But what I am concerned with here is a notion which is not a notion of epistemology but of metaphysics,

in some (I hope) nonpejorative sense. We ask whether something might have been true, or might have been false. Well, if something is false, it's obviously not necessarily true. If it is true, might it have been otherwise? Is it possible that, in this respect, the world should have been different from the way it is? If the answer is 'no', then this fact about the world is a necessary one. If the answer is 'yes', then this fact about the world is a contingent one. This in and of itself has nothing to do with anyone's knowledge of anything. It's certainly a philosophical thesis, and not a matter of obvious definitional equivalence, either that everything *a priori* is necessary or that everything necessary is *a priori*. Both concepts may be vague. That may be another problem. But at any rate they are dealing with two different domains, two different areas, the epistemological and the metaphysical. Consider, say, Fermat's last theorem—or the Goldbach conjecture. The Goldbach conjecture says that an even number greater than 2 must be the sum of two prime numbers. If this is true, it is presumably necessary, and, if it is false, presumably necessarily false. We are taking the classical view of mathematics here and assume that in mathematical reality it is either true or false.

If the Goldbach conjecture is false, then there is an even number, n, greater than 2, such that for no primes p_1 and p_2, both $< n$, does $n = p_1 + p_2$. This fact about n, if true, is verifiable by direct computation, and thus is necessary if the results of arithmetical computations are necessary. On the other hand, if the conjecture is true, then every even number exceeding 2 is the sum of two primes. Could it then be the case that, although in fact every such even number is the sum of two primes, there might have been such an even number which was not the sum of two primes? What would that mean? Such a number would have to be one of 4, 6, 8, 10, . . .; and, by hypothesis, since we are assuming Goldbach's conjecture to be true, each of these can be shown, again by direct computation, to be the sum of two primes. Goldbach's conjecture, then, cannot be contingently true or false; whatever truth-value it has belongs to it by necessity.

But what we can say, of course, is that right now, as far as we know, the question can come out either way. So, in the absence of a mathematical proof deciding this question, none of us has any *a priori* knowledge about this question in either direction. We don't know whether Goldbach's conjecture is true or false. So right now we certainly don't know anything *a priori* about it.

Perhaps it will be alleged that we *can* in principle know *a priori* whether it is true. Well, maybe we can. Of course an infinite mind which can search through all the numbers can or could. But I don't know whether a finite mind can or could. Maybe there just is no mathematical proof whatsoever which decides the conjecture. At any rate this might or might not be the case. Maybe there is a mathematical proof deciding this question; maybe every mathematical question is decidable by an intuitive proof or disproof. Hilbert thought so; others have thought not; still others have thought the question unintelligible unless the notion of intuitive proof is replaced by that of formal proof in a single system. Certainly no one formal system decides all mathematical questions, as we know from Gödel. At any rate, and this is the important thing, the question is not trivial; even though someone said that it's necessary, if true at all, that every even number is the sum of two primes, it doesn't follow that anyone knows anything *a priori* about it. It doesn't even seem to me to follow without some further philosophical argument (it is an interesting philosophical question) that anyone *could* know anything *a priori* about it. The 'could', as I said, involves some other modality. We mean that even if no one, perhaps even in the future, knows or will know *a priori* whether Goldbach's conjecture is right, in principle there is a way, which *could* have been

used, of answering the question *a priori*. This assertion is not trivial.

The terms 'necessary' and '*a priori*', then, as applied to statements, are *not* obvious synonyms. There may be a philosophical argument connecting them, perhaps even identifying them; but an argument is required, not simply the observation that the two terms are clearly interchangeable. (I will argue below that in fact they are not even coextensive—that necessary *a posteriori* truths, and probably contingent *a priori* truths, both exist.)

I think people have thought that these two things must mean the same for these reasons:

First, if something not only happens to be true in the actual world but is also true in all possible worlds, then, of course, just by running through all the possible worlds in our heads, we ought to be able with enough effort to see, if a statement is necessary, that it is necessary, and thus know it *a priori*. But really this is not so obviously feasible at all.

Second, I guess it's thought that, conversely, if something is known *a priori* it must be necessary, because it was known without looking at the world. If it depended on some contingent feature of the actual world, how could you know it without looking? Maybe the actual world is one of the possible worlds in which it would have been false. This depends on the thesis that there can't be a way of knowing about the actual world without looking that wouldn't be a way of knowing the same thing about every possible world. This involves problems of epistemology and the nature of knowledge; and of course it is very vague as stated. But it is not really *trivial* either. More important than any particular example of something which is alleged to be necessary and not *a priori* or *a priori* and not necessary, is to see that the notions are different, that it's not trivial to argue on the basis of something's being something which maybe we can only know *a posteriori*, that it's not

a necessary truth. It's not trivial, just because something is known in some sense *a priori*, that what is known is a necessary truth.

Another term used in philosophy is 'analytic'. Here it won't be too important to get any clearer about this in this talk. The common examples of analytic statements, nowadays, are like 'bachelors are unmarried'. Kant (someone just pointed out to me) gives as an example 'gold is a yellow metal', which seems to me an extraordinary one, because it's something I think that can turn out to be false. At any rate, let's just make it a matter of stipulation that an analytic statement is, in some sense, true by virtue of its meaning and true in all possible worlds by virtue of its meaning. Then something which is analytically true will be both necessary and *a priori*. (That's sort of stipulative.)

Another category I mentioned was that of certainty. Whatever certainty is, it's clearly not obviously the case that everything which is necessary is certain. Certainty is another epistemological notion. Something can be known, or at least rationally believed, *a priori*, without being quite certain. You've read a proof in the math book; and, though you think it's correct, maybe you've made a mistake. You often do make mistakes of this kind. You've made a computation, perhaps with an error.

There is one more question I want to go into in a preliminary way. Some philosophers have distinguished between essentialism, the belief in modality *de re*, and a mere advocacy of necessity, the belief in modality *de dicto*. Now, some people say: Let's *give* you the concept of necessity.[11] A much worse

[11] By the way, it's a common attitude in philosophy to think that one shouldn't introduce a notion until it's been rigorously defined (according to some popular notion of rigor). Here I am just dealing with an intuitive notion and will keep on the level of an intuitive notion. That is, we think that some things, though they are in fact the case, might have been otherwise. I might not have given these lectures today. If that's right, then it is *possible* that I

thing, something creating great additional problems, is whether we can say of any particular that it has necessary or contingent properties, even make the distinction between necessary and contingent properties. Look, it's only a *statement* or a *state of affairs* that can be either necessary or contingent! Whether a *particular* necessarily or contingently has a certain property depends on the way it's described. This is perhaps closely related to the view that the way we refer to particular things is by a description. What is Quine's famous example? If we consider the number 9, does it have the property of necessary oddness? Has that number got to be odd in all possible worlds? Certainly it's true in all possible worlds, let's say, it couldn't have been otherwise, that *nine* is odd. Of course, 9 could also be equally well picked out as *the number of planets*. It is *not* necessary, not true in all possible worlds, that the number of planets is odd. For example if there had been eight planets, the number of planets would not have been odd. And so it's thought: Was it necessary or contingent that Nixon won the election? (It might seem contingent, unless one has some view of some inexorable processes. . . .) But this is a contingent property of Nixon only relative to our referring to him as 'Nixon' (assuming 'Nixon' doesn't mean 'the man who won the election at such and such a time'). But if we designate Nixon as 'the man who won the election in 1968', then it will be a necessary truth, of course, that the man who won the election in 1968, won the election in 1968. Similarly, whether an object has the same property in all possible worlds depends

---

wouldn't have given these lectures today. Quite a different question is the epistemological question, how any particular person knows that I gave these lectures today. I suppose in that case he does know this is *a posteriori*. But, if someone were born with an innate belief that I was going to give these lectures today, who knows? Right now, anyway, let's suppose that people know this *a posteriori*. At any rate, the two questions being asked are different.

not just on the object itself, but on how it is described. So it's argued.

It is even suggested in the literature, that though a notion of necessity may have some sort of intuition behind it (we do think some things could have been otherwise; other things we don't think could have been otherwise), this notion [of a distinction between necessary and contingent properties] is just a doctrine made up by some bad philosopher, who (I guess) didn't realize that there are several ways of referring to the same thing. I don't know if some philosophers have not realized this; but at any rate it is very far from being true that this idea [that a property can meaningfully be held to be essential or accidental to an object independently of its description] is a notion which has no intuitive content, which means nothing to the ordinary man. Suppose that someone said, pointing to Nixon, 'That's the guy who might have lost'. Someone else says 'Oh no, if you describe him as "Nixon", then he might have lost; but, of course, describing him as the winner, then it is not true that he might have lost'. Now which one is being the philosopher, here, the unintuitive man? It seems to me obviously to be the second. The second man has a philosophical theory. The first man would say, and with great conviction, 'Well, of course, the winner of the election *might have been someone else*. The actual winner, had the course of the campaign been different, might have been the loser, and someone else the winner; or there might have been no election at all. So, such terms as "the winner" and "the loser" don't designate the same objects in all possible worlds. On the other hand, the term "Nixon" is just a *name* of *this man*'. When you ask whether it is necessary or contingent that *Nixon* won the election, you are asking the intuitive question whether in some counterfactual situation, *this man* would in fact have lost the election. If someone thinks that the notion of a necessary or contingent property (forget whether there *are* any nontrivial

necessary properties [and consider] just the *meaningfulness* of the notion[12]) is a philosopher's notion with no intuitive content, he is wrong. Of course, some philosophers think that something's having intuitive content is very inconclusive evidence in favor of it. I think it is very heavy evidence in favor of anything, myself. I really don't know, in a way, what more conclusive evidence one can have about anything, ultimately speaking. But, in any event, people who think the notion of accidental property unintuitive have intuition reversed, I think.

Why have they thought this? While there are many motivations for people thinking this, one is this: The question of essential properties so-called is supposed to be equivalent (and it is equivalent) to the question of 'identity across possible worlds'. Suppose we have someone, Nixon, and there's another possible world where there is no one with all the properties Nixon has in the actual world. Which one of these other people, if any, is Nixon? Surely you must give some criterion of identity here! If you have a criterion of identity, then you just look in the other possible worlds at the man who is Nixon; and the question whether, in that other possible world, Nixon has certain properties, is well defined. It is also supposed to be well defined, in terms of such notions, whether it's true in all possible worlds, or there are some possible worlds in which Nixon didn't win the election. But, it's said, the problems of giving such criteria of identity are very difficult. Sometimes

[12] The example I gave asserts a certain property—electoral victory—to be *accidental* to Nixon, independently of how he is described. Of course, if the notion of accidental property is meaningful, the notion of essential property must be meaningful also. This is not to say that there *are* any essential properties—though, in fact, I think there are. The usual argument questions the *meaningfulness* of essentialism, and says that whether a property is accidental or essential to an object depends on how it is described. It is thus *not* the view that all properties are accidental. Of course, it is also not the view, held by some idealists, that all properties are essential, all relations internal.

in the case of numbers it might seem easier (but even here it's argued that it's quite arbitrary). For example, one might say, and this is surely the truth, that if position in the series of numbers is what makes the number 9 what it is, then if (in another world) the number of planets had been 8, the number of planets would be a different number from the one it actually is. You wouldn't say that that number then is to be identified with our number 9 in this world. In the case of other types of objects, say people, material objects, things like that, has anyone given a set of necessary and sufficient conditions for identity across possible worlds?

Really, adequate necessary and sufficient conditions for identity which do not beg the question are very rare in any case. Mathematics is the only case I really know of where they are given even *within* a possible world, to tell the truth. I don't know of such conditions for identity of material objects over time, or for people. Everyone knows what a problem this is. But, let's forget about that. What seems to be more objectionable is that this depends on the wrong way of looking at what a possible world is. One thinks, in this picture, of a possible world as if it were like a foreign country. One looks upon it as an observer. Maybe Nixon has moved to the other country and maybe he hasn't, but one is given only qualities. One can observe all his qualities, but, of course, one doesn't observe that someone is Nixon. One observes that something has red hair (or green or yellow) but not whether something is Nixon. So we had better have a way of telling in terms of properties when we run into the same thing as we saw before; we had better have a way of telling, when we come across one of these other possible worlds, who was Nixon.

Some logicians in their formal treatment of modal logic may encourage this picture. A prominent example, perhaps, is myself. Nevertheless, intuitively speaking, it seems to me not

to be the right way of thinking about the possible worlds. A possible world isn't a distant country that we are coming across, or viewing through a telescope. Generally speaking, another possible world is too far away. Even if we travel faster than light, we won't get to it. A possible world is *given by the descriptive conditions we associate with it.* What do we mean when we say 'In some other possible world I would not have given this lecture today?' We just imagine the situation where I didn't decide to give this lecture or decided to give it on some other day. Of course, we don't imagine everything that is true or false, but only those things relevant to my giving the lecture; but, in theory, everything needs to be decided to make a total description of the world. We can't really imagine that except in part; that, then, is a 'possible world'. Why can't it be part of the *description* of a possible world that it contains *Nixon* and that in that world *Nixon* didn't win the election? It might be a question, of course, whether such a world *is* possible. (Here it would seem, *prima facie*, to be clearly possible.) But, once we see that such a situation is possible, then we are given that the man who might have lost the election or did lose the election in this possible world is Nixon, because that's part of the description of the world. 'Possible worlds' are *stipulated*, not *discovered* by powerful telescopes. There is no reason why we cannot *stipulate* that, in talking about what would have happened to Nixon in a certain counterfactual situation, we are talking about what would have happened to *him*.

Of course, if someone makes the demand that every possible world has to be described in a purely qualitative way, we can't say, 'Suppose Nixon had lost the election', we must say, instead, something like, 'Suppose a man with a dog named Checkers, who looks like a certain David Frye impersonation, is in a certain possible world and loses the election.' Well, does he resemble Nixon enough to be identified with Nixon? A very explicit and blatant example of this way of looking at

things is David Lewis's counterpart theory,[13] but the literature on quantified modality is replete with it.[14] Why need we make this demand? That is not the way we ordinarily think of counterfactual situations. We just say 'suppose this man had

[13] David K. Lewis, 'Counterpart Theory and Quantified Modal Logic', *Journal of Philosophy* **65** (1968), 113–126. Lewis's elegant paper also suffers from a purely formal difficulty: on his interpretation of quantified modality, the familiar law $(y)\,((x)A(x) \supset A(y))$ fails, if $A(x)$ is allowed to contain modal operators. (For example, $(\exists y)\,((x)\,\Diamond\,(x \neq y))$ is satisfiable but $(\exists y)\,\Diamond\,(y \neq y)$ is not.) Since Lewis's formal model follows rather naturally from his philosophical views on counterparts, and since the failure of universal instantiation for modal properties is intuitively bizarre, it seems to me that this failure constitutes an additional argument against the plausibility of his philosophical views. There are other, lesser, formal difficulties as well. I cannot elaborate here.

Strictly speaking, Lewis's view is not a view of 'transworld identification'. Rather, he thinks that similarities across possible worlds determine a counterpart relation which need be neither symmetric nor transitive. The counterpart of something in another possible world is *never* identical with the thing itself. Thus if we say 'Humphrey might have won the election (if only he had done such-and-such), we are not talking about something that might have happened to *Humphrey* but to someone else, a "counterpart".' Probably, however, Humphrey could not care less whether someone *else*, no matter how much resembling him, would have been victorious in another possible world. Thus, Lewis's view seems to me even more bizarre than the usual notions of transworld identification that it replaces. The important issues, however, are common to the two views: the supposition that other possible worlds are like other dimensions of a more inclusive universe, that they can be given only by purely qualitative descriptions, and that therefore either the identity relation or the counterpart relation must be established in terms of qualitative resemblance.

Many have pointed out to me that the father of counterpart theory is probably Leibnitz. I will not go into such a historical question here. It would also be interesting to compare Lewis's views with the Wheeler-Everett interpretation of quantum mechanics. I suspect that this view of physics may suffer from philosophical problems analogous to Lewis's counterpart theory; it is certainly very similar in spirit.

[14] Another *locus classicus* of the views I am criticizing, with more philosophical exposition than Lewis's paper, is a paper by David Kaplan on transworld identification. Unfortunately, this paper has never been published. It does not represent Kaplan's present position.

lost'. It is *given* that the possible world contains *this man*, and that in that world, he had lost. There may be a problem about what intuitions about possibility come to. But, if we have such an intuition about the possibility of *that* (*this man's* electoral loss), then it is about the possibility of *that*. It need not be identified with the possibility of a man looking like such and such, or holding such and such political views, or otherwise qualitatively described, having lost. We can point to the *man*, and ask what might have happened to *him*, had events been different.

It might be said 'Let's suppose that this is true. It comes down to the same thing, because whether Nixon could have had certain properties, different from the ones he actually has, is equivalent to the question whether the criteria of identity across possible worlds include that Nixon does not have these properties'. But it doesn't really come to the same thing, because the usual notion of a criterion of transworld identity demands that we give purely qualitative necessary and sufficient conditions for someone being Nixon. If we can't imagine a possible world in which Nixon doesn't have a certain property, then it's a necessary condition of someone being Nixon. Or a necessary property of Nixon that he [has] that property. For example, supposing Nixon is in fact a human being, it would seem that we cannot think of a possible counterfactual situation in which he was, say, an inanimate object; perhaps it is not even possible for him not to have been a human being. Then it will be a necessary fact about Nixon that in all possible worlds where he exists at all, he is human or anyway he is not an inanimate object. This has nothing to do with any requirement that there be purely qualitative *sufficient* conditions for Nixonhood which we can spell out. And should there be? Maybe there is some argument that there should be, but we can consider these questions about *necessary* conditions without going into any question about *sufficient* conditions.

Further, even if there were a purely qualitative set of necessary and sufficient conditions for being Nixon, the view I advocate would not demand that we find these conditions *before* we can ask whether Nixon might have won the election, nor does it demand that we restate the question in terms of such conditions. We can simply consider *Nixon* and ask what might have happened to *him* had various circumstances been different. So the two views, the two ways of looking at things, do seem to me to make a difference.

Notice this question, whether Nixon could not have been a human being, is a clear case where the question asked is not epistemological. Suppose Nixon actually turned out to be an automaton. That might happen. We might need evidence whether Nixon is a human being or an automaton. But that is a question about our knowledge. The question of whether Nixon might have not been a human being, given that he is one, is not a question about knowledge, *a posteriori* or *a priori*. It's a question about, even though such and such things are the case, what might have been the case otherwise.

This table is composed of molecules. Might it not have been composed of molecules? Certainly it was a scientific discovery of great moment that it was composed of molecules (or atoms). But could anything be this very object and not be composed of molecules? Certainly there is some feeling that the answer to that must be 'no'. At any rate, it's hard to imagine under what circumstances you would have this very object and find that it is not composed of molecules. A quite different question is whether it is in fact composed of molecules in the actual world and how we know this. (I will go into more detail about these questions about essence later on.)

I wish at this point to introduce something which I need in the methodology of discussing the theory of names that I'm talking about. We need the notion of 'identity across possible worlds' as it's usually and, as I think, somewhat misleadingly

called,[15] to explicate one distinction that I want to make now. What's the difference between asking whether it's necessary that 9 is greater than 7 or whether it's necessary that the number of planets is greater than 7? Why does one show anything more about essence than the other? The answer to this might be intuitively 'Well, look, the number of planets might have been different from what it in fact is. It doesn't make any sense, though, to say that nine might have been different from what it in fact is'. Let's use some terms quasi-technically. Let's call something a *rigid designator* if in every possible world it designates the same object, a *nonrigid* or *accidental designator* if that is not the case. Of course we don't require that the objects exist in all possible worlds. Certainly Nixon might not have existed if his parents had not gotten married, in the normal course of things. When we think of a property as essential to an object we usually mean that it is true of that object in any case where it would have existed. A rigid designator of a necessary existent can be called *strongly rigid.*

One of the intuitive theses I will maintain in these talks is that *names* are rigid designators. Certainly they seem to satisfy the intuitive test mentioned above: although someone other than the U.S. President in 1970 might have been the U.S. President in 1970 (e.g., Humphrey might have), no one other than Nixon might have been Nixon. In the same way, a

[15] Misleadingly, because the phrase suggests that there is a special problem of 'transworld identification', that we cannot trivially stipulate whom or what we are talking about when we imagine another possible world. The term 'possible world' may also mislead; perhaps it suggests the 'foreign country' picture. I have sometimes used 'counterfactual situation' in the text; Michael Slote has suggested that 'possible state (or history) of the world' might be less misleading than 'possible world'. It is better still, to avoid confusion, not to say, 'In some possible world, Humphrey would have won' but rather, simply, 'Humphrey might have won'. The apparatus of possible words has (I hope) been very useful as far as the set-theoretic model-theory of quantified modal logic is concerned, but has encouraged philosophical pseudo-problems and misleading pictures.

designator rigidly designates a certain object if it designates that object wherever the object exists; if, in addition, the object is a necessary existent, the designator can be called *strongly rigid.* For example, 'the President of the U.S. in 1970' designates a certain man, Nixon; but someone else (e.g., Humphrey) might have been the President in 1970, and Nixon might not have; so this designator is not rigid.

In these lectures, I will argue, intuitively, that proper names are rigid designators, for although the man (Nixon) might not have been the President, it is not the case that he might not have been Nixon (though he might not have been *called* 'Nixon'). Those who have argued that to make sense of the notion of rigid designator, we must antecedently make sense of 'criteria of transworld identity' have precisely reversed the cart and the horse; it is *because* we can refer (rigidly) to Nixon, and stipulate that we are speaking of what might have happened to *him* (under certain circumstances), that 'transworld identifications' are unproblematic in such cases.[16]

The tendency to demand purely qualitative descriptions of counterfactual situations has many sources. One, perhaps, is the confusion of the epistemological and the metaphysical, between a prioricity and necessity. If someone identifies necessity with a prioricity, and thinks that objects are named by means of uniquely identifying properties, he may think that it is the properties used to identify the object which, being known about it *a priori*, must be used to identify it in all possible worlds, to find out which object is Nixon. As against this, I repeat: (1) Generally, things aren't 'found out' about a counterfactual situation, they are stipulated; (2) possible worlds

[16] Of course I don't imply that language contains a name for every object. Demonstratives can be used as rigid designators, and free variables can be used as rigid designators of unspecified objects. Of course when we specify a counterfactual situation, we do not describe the whole possible world, but only the portion which interests us.

need not be given purely qualitatively, as if we were looking at them through a telescope. And we will see shortly that the properties an object has in every counterfactual world have nothing to do with properties used to identify it in the actual world.[17]

Does the 'problem' of 'transworld identification' make any sense? Is it *simply* a pseudo-problem? The following, it seems to me, can be said for it. Although the statement that England fought Germany in 1943 perhaps cannot be *reduced* to any statement about individuals, nevertheless in some sense it is not a fact 'over and above' the collection of all facts about persons, and their behavior over history. The sense in which facts about nations are not facts 'over and above' those about persons can be expressed in the observation that a description of the world mentioning all facts about persons but omitting those about nations can be a *complete* description of the world, from which the facts about nations follow. Similarly, perhaps, facts about material objects are not facts 'over and above' facts about their constituent molecules. We may then ask, given a description of a non-actualized possible situation in terms of people, whether England still exists in that situation, or whether a certain nation (described, say, as the one where Jones lives) which would exist in that situation, is England. Similarly, given certain counterfactual vicissitudes in the history of the molecules of a table, T, one may ask whether T would exist, in that situation, or whether a certain bunch of molecules, which in that situation would constitute a table, constitute the very same table T. In each case, we seek criteria of identity across possible worlds for certain particulars in terms of those for other, more 'basic', particulars. If statements about nations (or tribes) are not *reducible* to those about other more 'basic' constituents, if there is some 'open texture' in the relationship between them, we can hardly expect to give hard and fast identity criteria;

[17] See Lecture I, p. 53 (on Nixon), and Lecture II, pp. 74–7.

nevertheless, in concrete cases we may be able to answer whether a certain bunch of molecules would still constitute T, though in some cases the answer may be indeterminate. I think similar remarks apply to the problem of identity over time; here too we are usually concerned with determinacy, the identity of a 'complex' particular in terms of more 'basic' ones. (For example, if various parts of a table are replaced, is it the same object?[18])

Such a conception of 'transworld identification', however, differs considerably from the usual one. First, although we can try to describe the world in terms of molecules, there is no impropriety in describing it in terms of grosser entities: the statement that *this table* might have been placed in another room is perfectly proper, in and of itself. We *need* not use the description in terms of molecules, or even grosser parts of the table, though we *may*. Unless we assume that some particulars are 'ultimate', 'basic' particulars, no type of description need be regarded as privileged. We can ask whether *Nixon* might have lost the election without further subtlety, and usually no further subtlety is required. Second, it is not assumed that necessary and sufficient conditions for what kinds of collections

[18] There is some vagueness here. If a chip, or molecule, of a given table had been replaced by another one, we would be content to say that we have the same table. But if too many chips were different, we would seem to have a different one. The same problem can, of course, arise for identity over time.

Where the identity relation is vague, it may seem intransitive; a chain of apparent identities may yield an apparent non-identity. Some sort of 'counterpart' notion (though not with Lewis's philosophical underpinnings of resemblance, foreign country worlds, etc.), may have some utility here. One could say that strict identity applies only to the particulars (the molecules), and the counterpart relation to the particulars 'composed' of them, the tables. The counterpart relation can then be declared to be vague and intransitive. It seems, however, utopian to suppose that we will ever reach a level of ultimate, basic particulars for which identity relations are never vague and the danger of intransitivity is eliminated. The danger usually does not arise in practice, so we ordinarily can speak simply of identity without worry. Logicians have not developed a logic of vagueness.

of molecules make up this table are possible; this fact I just mentioned. Third, the attempted notion deals with criteria of identity of particulars in terms of other *particulars*, not qualities. I can refer to the table before me, and ask what might have happened to it under certain circumstances; I can also refer to its molecules. If, on the other hand, it is demanded that I describe each counterfactual situation purely qualitatively, then I can only ask whether *a table*, of such and such color, and so on, would have certain properties; whether the table in question would be *this table*, table *T*, is indeed moot, since all reference to objects, as opposed to qualities, has disappeared. It is often said that, if a counterfactual situation is described as one which would have happened to *Nixon*, and if it is not assumed that such a description is reducible to a purely qualitative one, then mysterious 'bare particulars' are assumed, propertyless substrata underlying the qualities. This is not so: I think that Nixon is a Republican, not merely that he lies in back of Republicanism, whatever that means; I also think he might have been a Democrat. The same holds for any other properties Nixon may possess, except that some of these properties may be essential. What I do deny is that a particular is nothing but a 'bundle of qualities', whatever that may mean. If a quality is an abstract object, a bundle of qualities is an object of an even higher degree of abstraction, not a particular. Philosophers have come to the opposite view through a false dilemma: they have asked, are these objects *behind* the bundle of qualities, or is the object *nothing but* the bundle? Neither is the case; this table is wooden, brown, in the room, etc. It has all these properties and is not a thing without properties, behind them; but it should not therefore be identified with the set, or 'bundle', of its properties, nor with the subset of its essential properties. Don't ask: how can I identify this table in another possible world, except by its properties? I have the table in my hands, I can point to it, and when I ask whether *it* might have been in

another room, I am talking, by definition, about *it*. I don't have to identify it after seeing it through a telescope. If I am talking about it, I am talking about *it*, in the same way as when I say that our hands might have been painted green, I have stipulated that I am talking about greenness. Some properties of an object may be essential to it, in that it could not have failed to have them. But these properties are not used to identify the object in another possible world, for such an identification is not needed. Nor need the essential properties of an object be the properties used to identify it in the actual world, if indeed it is identified in the actual world by means of properties (I have up to now left the question open).

So: the question of transworld identification makes *some* sense, in terms of asking about the identity of an object *via* questions about its component parts. But these parts are not qualities, and it is not an object resembling the given one which is in question. Theorists have often said that we identify objects across possible worlds as objects resembling the given one in the most important respects. On the contrary, Nixon, had he decided to act otherwise, might have avoided politics like the plague, though privately harboring radical opinions. Most important, even when we *can* replace questions about an object by questions about its parts, we *need* not do so. We can refer to the object and ask what might have happened to *it*. So, we do not begin with worlds (which are supposed somehow to be real, and whose qualities, but not whose objects, are perceptible to us), and then ask about criteria of transworld identification; on the contrary, we begin with the objects, which we *have*, and can identify, in the actual world. We can then ask whether certain things might have been true of the objects.

Above I said that the Frege-Russell view that names are introduced by description could be taken either as a theory of the meaning of names (Frege and Russell seemed to take it this

way) or merely as a theory of their reference. Let me give an example, not involving what would usually be called a 'proper name,' to illustrate this. Suppose someone stipulates that 100 degrees centigrade is to be the temperature at which water boils at sea level. This isn't completely precise because the pressure may vary at sea level. Of course, historically, a more precise definition was given later. But let's suppose that this were the definition. Another sort of example in the literature is that one meter is to be the length of S where S is a certain stick or bar in Paris. (Usually people who like to talk about these definitions then try to make 'the length of' into an 'operational' concept. But it's not important.)

Wittgenstein says something very puzzling about this. He says: 'There is one thing of which one can say neither that it is one meter long nor that it is not one meter long, and that is the standard meter in Paris. But this is, of course, not to ascribe any extraordinary property to it, but only to mark its peculiar role in the language game of measuring with a meter rule.'[19] This seems to be a very 'extraordinary property', actually, for any stick to have. I think he must be wrong. If the stick is a stick, for example, 39·37 inches long (I assume we have some different standard for inches), why isn't it one meter long? Anyway, let's suppose that he is wrong and that the stick is one meter long. Part of the problem which is bothering Wittgenstein is, of course, that this stick serves as a standard of length and so we can't attribute length to it. Be this as it may (well, it may not be), is the statement 'stick S is one meter long', a necessary truth? Of course its length might vary in time. We could make the definition more precise by stipulating that one meter is to be the length of S at a fixed time t_0. Is it then a necessary truth that stick S is one meter long at time t_0? Someone who thinks that everything one knows *a priori* is necessary might think: 'This is the *definition* of a meter. By

[19] *Philosophical Investigations*, § 50.

definition, stick S is one meter long at t_0. That's a necessary truth.' But there seems to me to be no reason so to conclude, even for a man who uses the stated definition of 'one meter'. For he's using this definition not to *give the meaning* of what he called the 'meter', but to *fix the reference*. (For such an abstract thing as a unit of length, the notion of reference may be unclear. But let's suppose it's clear enough for the present purposes.) He uses it to fix a reference. There is a certain length which he wants to mark out. He marks it out by an accidental property, namely that there is a stick of that length. Someone else might mark out the same reference by another accidental property. But in any case, even though he uses this to fix the reference of his standard of length, a meter, he can still say, 'if heat had been applied to this stick S at t_0, then at t_0 stick S would not have been one meter long.'

Well, why can he do this? Part of the reason may lie in some people's minds in the philosophy of science, which I don't want to go into here. But a simple answer to the question is this: Even if this is the *only* standard of length that he uses,[20] there is an intuitive difference between the phrase 'one meter' and the phrase 'the length of S at t_0'. The first phrase is meant to designate rigidly a certain length in all possible worlds, which in the actual world happens to be the length of the stick S at t_0. On the other hand 'the length of S at t_0' does not designate anything rigidly. In some counterfactual situations the stick might have been longer and in some shorter, if various stresses and strains had been applied to it. So we can say of this stick, the same way as we would of any other of the same substance and length, that if heat of a given quantity had been applied to it, it would have expanded to such and such a length. Such a

[20] Philosophers of science may see the key to the problem in a view that 'one meter' is a 'cluster concept'. I am asking the reader hypothetically to suppose that the 'definition' given is the *only* standard used to determine the metric system. I think the problem would still arise.

counterfactual statement, being true of other sticks with identical physical properties, will also be true of this stick. There is no conflict between that counterfactual statement and the definition of 'one meter' as 'the length of S at t_0', because the 'definition', properly interpreted, does *not* say that the phrase 'one meter' is to be *synonymous* (even when talking about counterfactual situations) with the phrase 'the length of S at t_0', but rather that we have *determined the reference* of the phrase 'one meter' by stipulating that 'one meter' is to be a *rigid* designator of the length which is in fact the length of S at t_0. So this does *not* make it a necessary truth that S is one meter long at t_0. In fact, under certain circumstances, S would not have been one meter long. The reason is that one designator ('one meter') is rigid and the other designator ('the length of S at t_0') is not.

What then, is the *epistemological* status of the statement 'Stick S is one meter long at t_0', for someone who has fixed the metric system by reference to stick S? It would seem that he knows it *a priori*. For if he used stick S to fix the reference of the term 'one meter', then as a result of this kind of 'definition' (which is not an abbreviative or synonymous definition), he knows automatically, without further investigation, that S is one meter long.[21] On the other hand, even if S is used as the standard of a meter, the *metaphysical* status of 'S is one meter long' will be that of a contingent statement, provided that 'one meter' is regarded as a rigid designator: under appropriate stresses and strains, heatings or coolings, S would have had a length other than one meter even at t_0. (Such statements as 'Water boils at 100°C at sea level' can have a similar status.) So in this sense, there are contingent *a priori* truths. More important for present purposes, though, than accepting this

[21] Since the truth he knows is contingent, I choose *not* to call it 'analytic', stipulatively requiring analytic truths to be both necessary and *a priori*. See footnote 63.

example as an instance of the contingent *a priori*, is its illustration of the distinction between 'definitions' which fix a reference and those which give a synonym.

In the case of names one might make this distinction too. Suppose the reference of a name is given by a description or a cluster of descriptions. If the name *means the same* as that description or cluster of descriptions, it will not be a rigid designator. It will not necessarily designate the same object in all possible worlds, since other objects might have had the given properties in other possible worlds, unless (of course) we happened to use essential properties in our description. So suppose we say, 'Aristotle is the greatest man who studied with Plato'. If we used that as a *definition*, the name 'Aristotle' is to mean 'the greatest man who studied with Plato'. Then of course in some other possible world that man might not have studied with Plato and some other man would have been Aristotle. If, on the other hand, we merely use the description to *fix the referent* then that man will be the referent of 'Aristotle' in all possible worlds. The only use of the description will have been to pick out to which man we mean to refer. But then, when we say counterfactually 'suppose Aristotle had never gone into philosophy at all', we need not mean 'suppose a man who studied with Plato, and taught Alexander the Great, and wrote this and that, and so on, had never gone into philosophy at all', which might seem like a contradiction. We need only mean, 'suppose that *that man* had never gone into philosophy at all'.

It seems plausible to suppose that, in some cases, the reference of a name is indeed fixed *via* a description in the same way that the metric system was fixed. When the mythical agent first saw Hesperus, he may well have fixed his reference by saying, 'I shall use "Hesperus" as a name of the heavenly body appearing in yonder position in the sky.' He then fixed the reference of 'Hesperus' by its apparent celestial position. Does it follow

that it is part of the *meaning* of the name that Hesperus has such and such position at the time in question? Surely not: if Hesperus had been hit earlier by a comet, it might have been visible at a different position at that time. In such a counterfactual situation we would say that Hesperus would not have occupied that position, but not that Hesperus would not have been Hesperus. The reason is that 'Hesperus' rigidly designates a certain heavenly body and 'the body in yonder position' does not—a different body, or no body might have been in that position, but no other body might have been Hesperus (though another body, not Hesperus, might have been *called* 'Hesperus'). Indeed, as I have said, I will hold that names are always rigid designators.

Frege and Russell certainly seem to have the full blown theory according to which a proper name is not a rigid designator and is synonymous with the description which replaced it. But another theory might be that this description is used to determine a rigid reference. These two alternatives will have different consequences for the questions I was asking before. If 'Moses' *means* 'the man who did such and such', then, if no one did such and such, Moses didn't exist; and maybe 'no one did such and such' is even an *analysis* of 'Moses didn't exist'. But if the description is used to fix a reference rigidly, then it's clear that that is *not* what is meant by 'Moses didn't exist', because we can ask, if we speak of a counterfactual case where no one did indeed do such and such, say, lead the Israelites out of Egypt, does it follow that, in such a situation, Moses wouldn't have existed? It would seem not. For surely Moses might have just decided to spend his days more pleasantly in the Egyptian courts. He might never have gone into either politics or religion at all; and in that case maybe no one would have done any of the things that the Bible relates of Moses. That doesn't in itself mean that in such a possible world Moses wouldn't have existed. If so, then

'Moses exists' means something different from 'the existence and uniqueness conditions for a certain description are fulfilled'; and therefore this does not give an analysis of the singular existential statement after all. If you give up the idea that this is a theory of meaning and make it into a theory of reference in the way that I have described, you give up some of the advantages of the theory. Singular existential statements and identity statements between names need some other analysis.

Frege should be criticized for using the term 'sense' in two senses. For he takes the sense of a designator to be its meaning; and he also takes it to be the way its reference is determined. Identifying the two, he supposes that both are given by definite descriptions. Ultimately, I will reject this second supposition too; but even were it right, I reject the first. A description may be used as synonymous with a designator, or it may be used to fix its reference. The two Fregean senses of 'sense' correspond to two senses of 'definition' in ordinary parlance. They should carefully be distinguished.[22]

[22] Usually the Fregean sense is now interpreted as the meaning, which must be carefully distinguished from a 'reference fixer'. We shall see below that for most speakers, unless they are the ones who initially give an object its name, the referent of the name is determined by a 'causal' chain of communication rather than a description.

In the formal semantics of modal logic, the 'sense' of a term t is usually taken to be the (possibly partial) function which assigns to each possible world H the referent of t in H. For a rigid designator, such a function is constant. This notion of 'sense' relates to that of 'giving a meaning', not that of fixing a reference. In this use of 'sense', 'one meter' has a constant function as its sense, though its reference is fixed by 'the length of S', which does not have a constant function as its sense.

Some philosophers have thought that descriptions, in English, are ambiguous, that sometimes they non-rigidly designate, in each world, the object (if any) satisfying the description, while sometimes they *rigidly* designate the object actually satisfying the description. (Others, inspired by Donnellan, say the description sometimes rigidly designates the object thought or presupposed to satisfy the description.) I find any such alleged ambiguities dubious. I know

I hope the idea of fixing the reference as opposed to actually defining one term as meaning the other is somewhat clear. There is really not enough time to go into everything in great detail. I think, even in cases where the notion of rigidity versus accidentality of designation cannot be used to make out the difference in question, some things called definitions really intend to fix a reference rather than to give the meaning of a phrase, to give a synonym. Let me give an example. π is supposed to be the ratio of the circumference of a circle to its diameter. Now, it's something that I have nothing but a vague intuitive feeling to argue for: It seems to me that here this Greek letter is not being used as *short for* the phrase 'the ratio of the circumference of a circle to its diameter' nor is it even used as short for a cluster of alternative definitions of π, whatever that might mean. It is used as a *name* for a real number, which in this case is necessarily the ratio of the circumference of a circle to its diameter. Note that here both 'π' and 'the ratio of the circumference of a circle to its diameter' are rigid designators, so the arguments given in the metric case are inapplicable. (Well, if someone doesn't see this, or thinks it's wrong, it doesn't matter.)

Let me return to the question about names which I raised. As I said, there is a popular modern substitute for the theory of Frege and Russell; it is adopted even by such a strong critic of many views of Frege and Russell, especially the latter, as

of no clear evidence for them which cannot be handled either by Russell's notion of scope or by the considerations alluded to in footnote 3, p. 25.

If the ambiguity does exist, then in the supposed *rigid* sense of 'the length of *S*', 'one meter' and 'the length of *S*' designate the same thing in all possible worlds and have the same (functional) 'sense'.

In the formal semantics of intensional logic, suppose we take a definite description to designate, in each world, the object satisfying the description. It is indeed useful to have an operator which transforms each description into a term which rigidly designates the object *actually* satisfying the description. David Kaplan has proposed such an operator and calls it 'Dthat'.

Strawson.[23] The substitute is that, although a name is not a disguised description it either abbreviates, or anyway its reference is determined by, some cluster of descriptions. The question is whether this is true. As I also said, there are stronger and weaker versions of this. The stronger version would say that the name is simply *defined*, synonymously, as the cluster of descriptions. It will then be necessary, not that Moses had any particular property in this cluster, but that he had the disjunction of them. There couldn't be any counterfactual situation in which he didn't do any of those things. I think it's clear that this is very implausible. People *have* said it—or maybe they haven't been intending to say that, but were using 'necessary' in some other sense. At any rate, for example, in Searle's article on proper names:

> To put the same point differently, suppose we ask, 'why do we have proper names at all?' Obviously to refer to individuals. 'Yes but descriptions could do that for us'. But only at the cost of specifying identity conditions every time reference is made: Suppose we agree to drop 'Aristotle' and use, say, 'the teacher of Alexander', then it is a necessary truth that the man referred to is Alexander's teacher—but it is a contingent fact that Aristotle ever went into pedagogy (though I am suggesting that it is a necessary fact that Aristotle has the logical sum, inclusive disjunction, of properties commonly attributed to him).[24]

Such a suggestion, if 'necessary' is used in the way I have been using it in this lecture, must clearly be false. (Unless he's got some very interesting essential property commonly attributed to Aristotle.) Most of the things commonly attributed to Aristotle are things that Aristotle might not have done at all. In a situation in which he didn't do them, we would describe that as a situation in which *Aristotle* didn't do them. This is not a distinction of scope, as happens sometimes in the case of

[23] P. F. Strawson, *Individuals*, Methuen, London, 1959, Ch. 6.

[24] Searle, op. cit. in Caton, *Philosophy and Ordinary Language*, p. 160.

descriptions, where someone might say that the man who taught Alexander might not have taught Alexander; though it could not have been true that: the man who taught Alexander didn't teach Alexander. This is Russell's distinction of scope. (I won't go into it.) It seems to me clear that this is not the case here. Not only is it true *of* the man Aristotle that he might not have gone into pedagogy; it is also true that we use the term 'Aristotle' in such a way that, in thinking of a counterfactual situation in which Aristotle didn't go into any of the fields and do any of the achievements we commonly attribute to him, still we would say that was a situation in which *Aristotle* did not do these things.[25] Well there are some things like the date, the period he lived in, that might be more imagined as necessary. Maybe those are things we commonly attribute to him. There are exceptions. Maybe it's hard to imagine how he could have lived 500 years later than he in fact did. That certainly raises at least a problem. But take a man who doesn't have any idea of the date. Many people just have some vague cluster of his most famous achievements. Not only each of these singly, but the possession of the entire disjunction of these properties, is just a contingent fact about Aristotle; and the statement

[25] The facts that 'the teacher of Alexander' is capable of scope distinctions in modal contexts and that it is not a rigid designator are both illustrated when one observes that the teacher of Alexander might not have taught Alexander (and, in such circumstances, would not have been the teacher of Alexander). On the other hand, it is not true that Aristotle might not have been Aristotle, although Aristotle might not have been *called* 'Aristotle', just as 2×2 might not have been *called* 'four'. (Sloppy, colloquial speech, which often confuses use and mention, may, of course, express the fact that someone might have been called, or not have been called, 'Aristotle' by saying that he might have been, or not have been, Aristotle. Occasionally, I have heard such loose usages adduced as counterexamples to the applicability of the present theory to ordinary language. Colloquialisms like these seem to me to create as little problem for my theses as the success of the 'Impossible Missions Force' creates for the modal law that the impossible does not happen.) Further, although under certain circumstances Aristotle would not have taught Alexander, these are not circumstances under which he would not have been Aristotle.

that Aristotle had this disjunction of properties is a contingent truth.

A man might know it *a priori* in some sense, if he in fact fixes the reference of 'Aristotle' as the man who did one of these things. Still it won't be a necessary truth for him. So this sort of example would be an example where a prioricity would not necessarily imply necessity, if the cluster theory of names were right. The case of fixing the reference of 'one meter' is a very clear example in which someone, just because he fixed the reference in this way, can in some sense know *a priori* that the length of this stick is a meter without regarding it as a necessary truth. Maybe the thesis about a prioricity implying necessity can be modified. It does appear to state some insight which might be important, and true, about epistemology. In a way an example like this may seem like a trivial counterexample which is not really the point of what some people think when they think that only necessary truths can be known *a priori*. Well, if the thesis that all *a priori* truth is necessary is to be immune from this sort of counterexample, it needs to be modified in some way. Unmodified it leads to confusion about the nature of reference. And I myself have no idea how it should be modified or restated, or if such a modification or restatement is possible.[26]

[26] If someone fixes a meter as 'the length of stick S at t_0', then in some sense he knows *a priori* that the length of stick S at t_0 is one meter, even though he uses this statement to express a contingent truth. But, merely by fixing a system of measurement, has he thereby *learned* some (contingent) *information* about the world, some new *fact* that he did not know before? It seems plausible that in some sense he did not, even though it is undeniably a contingent fact that S is one meter long. So there may be a case for reformulating the thesis that everything *a priori* is necessary so as to save it from this type of counterexample. As I said, I don't know how such a reformulation would go; the reformulation should not be such as to make the thesis trivial (e.g., by defining *a priori* as known to be *necessary* (instead of true) independently of experience); and the converse thesis would still be false.

Since I will not attempt such a reformulation, I shall consistently use the

Let me state then what the cluster concept theory of names is. (It really is a nice theory. The only defect I think it has is probably common to all philosophical theories. It's wrong. You may suspect me of proposing another theory in its place; but I hope not, because I'm sure it's wrong too if it is a theory.) The theory in question can be broken down into a number of theses, with some subsidiary theses if you want to see how it handles the problem of existence statements, identity statements, and so on. There are more theses if you take it in the stronger version as a theory of meaning. The speaker is *A*.

(1) To every name or designating expression '*X*', there corresponds a cluster of properties, namely the family of properties φ such that *A* believes 'φX'.

This thesis is true, because it can just be a definition. Now, of course, some people might think that not everything the speaker believes about *X* has anything to do with determining the reference of '*X*'. They might only be interested in a subset. But we can handle this later on by modifying some of the other theses. So this thesis is correct, by definition. The theses that follow, however, are all, I think, false.

(2) One of the properties, or some conjointly, are believed by *A* to pick out some individual uniquely.

This doesn't say that they do pick out something uniquely, just that *A* believes that they do. Another thesis is that he is correct.

(3) If most, or a weighted most, of the φ's are satisfied by one unique object *y*, then *y* is the referent of '*X*'.

Well, the theory says that the referent of '*X*' is supposed to be the thing satisfying, if not all the properties, 'enough' of them.

term '*a priori*' in the text so as to make statements whose truth follows from a reference-fixing 'definition' *a priori*.

Obviously *A* could be wrong about some things about *X*. You take some sort of a vote. Now the question is whether this vote should be democratic or have some inequalities among the properties. It seems more plausible that there should be some weighting, that some properties are more important than others. A theory really has to specify how this weighting goes. I believe that Strawson, to my surprise, explicitly states that democracy should rule here, so the most trivial properties are of equal weight with the most crucial.[27] Surely it is more plausible to suppose that there is some weighting. Let's say democracy doesn't necessarily rule. If there is any property that's completely irrelevant to the reference we can disenfranchise it altogether, by giving it weight 0. The properties can be regarded as members of a corporation. Some have more stock than others; some may even have only non-voting stock.

(4) If the vote yields no unique object, '*X*' does not refer.

(5) The statement, 'If *X* exists, then *X* has most of the φ's' is known *a priori* by the speaker.

(6) The statement, 'If *X* exists, then *X* has most of the φ's' expresses a necessary truth (in the idiolect of the speaker).

(6) need not be a thesis of the theory if someone doesn't think that the cluster is part of the meaning of the name. He could think that though he determines the reference of 'Aristotle' as the man who had most of the φ's, still there are certainly possible situations in which Aristotle wouldn't have had most of the φ's.

As I indicated, there are some subsidiary theses, though I won't go into them in detail. These would give the analyses of singular existential statements like, '"Moses exists" means

[27] Strawson, op. cit., pp. 191–92. Strawson actually considers the case of several speakers, pools their properties, and takes a democratic (equally weighted) vote. He requires only a sufficiently plurality, not a majority.

"enough of the properties φ are satisfied" '. Even the man who doesn't use the theory as a theory of meaning has some of these theses. For example, subsidiary to thesis 4, we should say that it is *a priori* true for the speaker that, if not enough of the φ's are satisfied, then *X* does not exist. Only if he holds the view as a theory of meaning, rather than of reference, would it also be *necessarily* true that, if not enough of the φ's are satisfied, *X* does not exist. In any case it will be something he knows *a priori*. (At least he will know it *a priori* provided he knows the proper theory of names.) Then there is also an analysis of identity statements along the same lines.

The question is, are any of these true? If true, they give a nice picture of what's going on. Preliminary to discussing these theses, let me mention that, often, when people specify which properties φ are relevant, they seem to specify them wrongly. That's just an incidental defect, though it is closely related to the arguments against the theory that I will give presently. Consider the example from Wittgenstein. What does he say the relevant properties are? 'When one says "Moses does not exist", this may mean various things. It may mean: the Israelites did not have a *single* leader when they withdrew from Egypt—or: their leader was not called Moses—or: there cannot have been anyone who accomplished all that the Bible relates of Moses. . . .' The gist of all this is that we know *a priori* that, if the Biblical story is substantially false, Moses did not exist. I have already argued that the Biblical story does not give *necessary* properties of Moses, that he might have lived without doing any of these things. Here I ask whether we know *a priori* that if Moses existed, he in fact did some or most of them. Is this really the cluster of properties that we should use here? Surely there is a distinction which is neglected in these kinds of remarks. The Biblical story might have been a complete legend, or it might have been a substantially false account of a real person. In the latter case, it seems to me that a scholar

could say that he supposes that, though Moses did exist, the things said of him in the Bible are substantially false. Such things occur in this very field of scholarship. Suppose that someone says that no prophet ever was swallowed by a big fish or a whale. Does it follow, on that basis, that Jonah did not exist? There still seems to be the question whether the Biblical account is a legendary account of no person or a legendary account built on a real person. In the latter case, it's only natural to say that, though Jonah did exist, no one did the things commonly related to him. I choose this case because while Biblical scholars generally hold that Jonah did exist, the account not only of his being swallowed by a big fish but even going to Nineveh to preach or anything else that is said in the Biblical story is assumed to be substantially false. But nevertheless there are reasons for thinking this was about a real prophet. If I had a suitable book along with me I could start quoting out of it: 'Jonah, the son of Amittai, was a real prophet, however such and such and such'. There are independent reasons for thinking this was not a pure legend about an imaginary character but one about a real character.[28]

[28] See, for example, H. L. Ginsberg, *The Five Megilloth and Jonah*, The Jewish Publication Society of America, 1969, p. 114: 'The "hero" of this tale, the prophet Jonah the son of Amittai, is a historical personage . . . (but) this book is not history but fiction.' The scholarly consensus regards all details about Jonah in the book as legendary and not even based on a factual substratum, excepting the bare statement that he was a Hebrew prophet, which is hardly uniquely identifying. Nor need he have been *called* 'Jonah' by the Hebrews; the '*J*' sound does not exist in Hebrew, and Jonah's historical existence is independent of whether we know his original Hebrew name or not. The fact that *we* call him Jonah cannot be used to single him out without circularity. The evidence for the historicity of Jonah comes from an independent reference to him in *II Kings*; but such evidence could have been available in the absence of any such other references—e.g., evidence that all Hebrew legends were about actual personages. Further, the statement that Jonah is a legend about a real person might have been *true*, even if there were no evidence for it. One may say, 'The Jonah of the book never existed,' as one may say, 'The Hitler of Nazi propaganda never existed.' As the quotation above shows, this usage

These examples could be modified. Maybe all we believe is that *the Bible relates of him* that such and such. This gives us another problem, because how do we know whom the Bible is referring to? The question of our reference is thrown back to the question of reference in the Bible. This leads to a condition which we ought to put in explicitly.

(C) For any successful theory, the account must not be circular. The properties which are used in the vote must not themselves involve the notion of reference in a way that it is ultimately impossible to eliminate.

Let me give an example where the noncircularity condition is clearly violated. The following theory of proper names is due to William Kneale in an article called 'Modality, De Dicto and De Re'.[29] It contains, I think, a clear violation of noncircularity conditions.

Ordinary proper names of people are not, as John Stuart Mill supposed, signs without sense. While it may be informative to tell a man that the most famous Greek philosopher was called Socrates, it is obviously trifling to tell him that Socrates was called Socrates; and the reason is simply that he cannot understand your use of the word 'Socrates' at the beginning of your statement unless he already knows that 'Socrates' means 'The individual called "Socrates" '.[30]

Here we have a theory of the reference of proper names. 'Socrates' just means 'the man called "Socrates".' Actually, of course, maybe not just one man can be called 'Socrates', and

need not coincide with the historian's view of whether Jonah ever existed. Ginsberg is writing for the lay reader, who, he assumes, will find his statement intelligible.

[29] In Ernest Nagel, Patrick Suppes, and Alfred Tarski, *Logic, Methodology and the Philosophy of Science: Proceedings of the 1960 International Congress*, Stanford University Press, 1962, 622–33.

[30] *Loc. cit.*, pp. 629–30.

some may call him 'Socrates' while others may not. Certainly that is a condition which under some circumstances is uniquely satisfied. Maybe only one man was called 'Socrates' by me on a certain occasion.

Kneale says it's trifling to tell someone that Socrates *was* called 'Socrates'. That isn't trifling on any view. Maybe the Greeks didn't call him 'Socrates'. Let's say that Socrates is called 'Socrates' by us—by *me* anyway. Suppose that's trifling. (I find it surprising that Kneale uses the past tense here; it is dubious that the Greeks *did* call him 'Socrates'—at least, the Greek name is pronounced differently. I will check the accuracy of the quotation for the next lecture.)

Kneale gives an argument for this theory. 'Socrates' must be analyzed as 'the individual called "Socrates" ', because how else can we explain the fact that it is trifling to be told that Socrates is called 'Socrates'? In some cases that's rather trifling. In the same sense, I suppose, you could get a good theory of the meaning of any expression in English and construct a dictionary. For example, though it may be informative to tell someone that horses are used in races, it is trifling to tell him that horses are called 'horses'. Therefore this could only be the case because the term 'horse', means in English 'the things called "horses" '. Similarly with any other expression which might be used in English. Since it's trifling to be told that sages are called 'sages', 'sages' just means 'the people called "sages" '. Now plainly this isn't really a very good argument, nor can it therefore be the only explanation of why it's trifling to be told that Socrates is called 'Socrates'. Let's not go into exactly why it's trifling. Of course, anyone who knows the use of 'is called' in English, even without knowing what the statement means, knows that if 'quarks' means something then 'quarks are called "quarks" ' will express a truth. He may not know what truth it expresses, because he doesn't know what a quark is. But his knowledge that it expresses a truth

does not have much to do with the meaning of the term 'quarks'.

We could go into this actually at great length. There are interesting problems coming out of this sort of passage. But the main reason I wanted to introduce it here is that as a theory of reference it would give a clear violation of the noncircularity condition. Someone uses the name 'Socrates'. How are we supposed to know to whom he refers? By using the description which gives the sense of it. According to Kneale, the description is 'the man called "Socrates" '. And here, (presumably, since this is supposed to be so trifling!) it tells us nothing at all. Taking it in this way it seems to be no theory of reference at all. We ask, 'To whom does he refer by "Socrates"?' And then the answer is given, 'Well, he refers to the man to whom he refers.' If this were all there was to the meaning of a proper name, then no reference would get off the ground at all.

So there's a condition to be satisfied; in the case of this particular theory it's obviously unsatisfied. The paradigm, amazingly enough, is even sometimes used by Russell as the descriptive sense, namely: 'the man called "Walter Scott" '. Obviously if the only descriptive senses of names we can think of are of the form 'the man called such and such', 'the man called "Walter Scott" ', 'the man called "Socrates" ', then whatever this relation of *calling* is is really what determines the reference and not any description like 'the man called "Socrates" '.

LECTURE II: JANUARY 22, 1970

Last time we ended up talking about a theory of naming which is given by a number of theses here on the board.

(1) To every name or designating expression 'X', there corresponds a cluster of properties, namely the family of those properties φ such that A believes 'φX'.

(2) One of the properties, or some conjointly, are believed by A to pick out some individual uniquely.

(3) If most, or a weighted most, of the φ's are satisfied by one unique object y, then y is the referent of 'X'.

(4) If the vote yields no unique object, 'X' does not refer.

(5) The statement, 'If X exists, then X has most of the φ's' is known *a priori* by the speaker.

(6) The statement, 'If X exists, then X has most of the φ's' expresses a necessary truth (in the idiolect of the speaker).

(C) For any successful theory, the account must not be circular. The properties which are used in the vote must not themselves involve the notion of reference in such a way that it is ultimately impossible to eliminate.

(C) is not a thesis but a condition on the satisfaction of the other theses. In other words, Theses (1)–(6) cannot be satisfied in a way which leads to a circle, in a way which does not lead to any independent determination of reference. The example I

gave last time of a blatantly circular attempt to satisfy these conditions was a theory of names mentioned by William Kneale. I was a little surprised at the statement of the theory when I was reading what I had copied down, so I looked it up again. I looked it up in the book to see if I'd copied it down accurately. Kneale *did* use the past tense. He said that though it is not trifling to be told that Socrates was the greatest philosopher of ancient Greece, it is trifling to be told that Socrates was called 'Socrates'. Therefore, he concludes, the name 'Socrates' must simply mean 'the individual called "Socrates" '. Russell, as I've said, in some places gives a similar analysis. Anyway, as stated using the past tense, the condition wouldn't be circular, because one certainly could decide to use the term 'Socrates' to refer to whoever was called 'Socrates' by the Greeks. But, of course, in that sense it's not at all trifling to be told that Socrates was called 'Socrates'. If this is any kind of fact, it might be false. Perhaps we know that *we* call him 'Socrates'; that hardly shows that the Greeks did so. In fact, of course, they may have pronounced the name differently. It may be, in the case of this particular name, that transliteration from the Greek is so good that the English version is not pronounced *very* differently from the Greek. But that won't be so in the general case. Certainly it is not trifling to be told that Isaiah was called 'Isaiah'. In fact, it is false to be told that Isaiah was called 'Isaiah'; the prophet wouldn't have recognized this name at all. And of course the Greeks didn't call their country anything like 'Greece'. Suppose we amend the thesis so that it reads: it's trifling to be told that Socrates is called 'Socrates' by us, or at least, by me, the speaker. Then in some sense this is fairly trifling. I don't think it is necessary or analytic. In the same way, it is trifling to be told that horses are called 'horses', without this leading to the conclusion that the word 'horse' simply *means* 'the animal called a "horse" '. As a theory of the reference of the name 'Socrates' it will lead immediately

to a vicious circle. If one was determining the referent of a name like 'Glunk' to himself and made the following decision, 'I shall use the term "Glunk" to refer to the man that I call "Glunk" ', this would get one nowhere. One had better have some independent determination of the referent of 'Glunk'. This is a good example of a blatantly circular determination. Actually sentences like 'Socrates is called "Socrates" ' are very interesting and one can spend, strange as it may seem, hours talking about their analysis. I actually did, once, do that. I won't do that, however, on this occasion. (See how high the seas of language can rise. And at the lowest points too.) Anyway this is a useful example of a violation of the noncircularity condition. The theory will satisfy all of these statements, perhaps, but it satisfies them only because there is some independent way of determining the reference independently of the particular condition: being the man called 'Socrates'.

I have already talked about, in the last lecture, Thesis (6). Theses (5) and (6), by the way, have converses. What I said for Thesis (5) is that the statement that if X exists, X has most of the φ's, is *a priori* true for the speaker. It will also be true under the given theory that certain converses of this statement hold true also *a priori* for the speaker, namely: if any unique thing has most of the properties φ in the properly weighted sense, it is X. Similarly a certain converse to this will be *necessarily* true, namely: if anything has most of the properties φ in the properly weighted sense, it is X. So really one can say that it is both *a priori* and necessary that something is X if and only if it uniquely has most of the properties φ. This really comes from the previous Theses (1)–(4), I suppose. And (5) and (6) really just say that a sufficiently reflective speaker grasps this theory of proper names. Knowing this, he therefore sees that (5) and (6) are true. The objections to Theses (5) and (6) will *not* be that some speakers are unaware of this theory and therefore don't know these things.

What I talked about in the last lecture is Thesis (6). It's been observed by many philosophers that, if the cluster of properties associated with a proper name is taken in a very narrow sense, so that only one property is given any weight at all, let's say one definite description to pick out the referent—for example, Aristotle was the philosopher who taught Alexander the Great—then certain things will seem to turn out to be necessary truths which are not necessary truths—in this case, for example, that Aristotle taught Alexander the Great. But as Searle said, it is not a necessary truth but a contingent one that Aristotle ever went into pedagogy. Therefore, he concludes that one must drop the original paradigm of a single description and turn to that of a cluster of descriptions.

To summarize some things that I argued last time, this is not the correct answer (whatever it may be) to this problem about necessity. For Searle goes on to say,

> Suppose we agree to drop 'Aristotle' and use, say, 'the teacher of Alexander', then it is a necessary truth that the man referred to is Alexander's teacher—but it is a contingent fact that Aristotle ever went into pedagogy, though I am suggesting that it is a necessary fact that Aristotle has the logical sum, inclusive disjunction, of properties commonly attributed to him. . . .[31]

This is what is not so. It just is not, in any intuitive sense of necessity, a necessary truth that Aristotle had the properties commonly attributed to him. There is a certain theory, perhaps popular in some views of the philosophy of history, which might both be deterministic and yet at the same time assign a great role to the individual in history. Perhaps Carlyle would associate with the meaning of the name of a great man his achievements. According to such a view it will be necessary, once a certain individual is born, that he is destined to perform

[31] Searle, 'Proper Names', in Caton, op. cit., p. 160.

various great tasks and so it will be part of the very nature of Aristotle that he should have produced ideas which had a great influence on the western world. Whatever the merits of such a view may be as a view of history or the nature of great men, it does not seem that it should be trivially true on the basis of a theory of proper names. It would seem that it's a contingent fact that Aristotle ever did *any* of the things commonly attributed to him today, *any* of these great achievements that we so much admire. I must say that there is *something* to this feeling of Searle's. When I hear the name 'Hitler', I do get an illusory 'gut feeling' that it's sort of analytic that that man was evil. But really, probably not. Hitler might have spent all his days in quiet in Linz. In that case we would not say that then this man would not have been Hitler, for we use the name 'Hitler' just as the name of that man, even in describing other possible worlds. (This is the notion which I called a *rigid designator* in the previous talk.) Suppose we do decide to pick out the reference of 'Hitler', as the man who succeeded in having more Jews killed than anyone else managed to do in history. That is the way we pick out the reference of the name; but in another counterfactual situation where some one else would have gained this discrédit, we wouldn't say that in that case that other man would have been Hitler. If Hitler had never come to power, Hitler would not have had the property which I am supposing we use to fix the reference of his name. Similarly, even if we define what a meter is by reference to the standard meter stick, it will be a contingent truth and not a necessary one that that particular stick is one meter long. If it had been stretched, it would have been longer than one meter. And that is because we use the term 'one meter' rigidly to designate a certain length. Even though we fix what length we are designating by an accidental property of that length, just as in the case of the name of the man we may pick the man out by an accidental property of the man, still we use the name

to designate that man or that length in all possible worlds. The property we use need not be one which is regarded in any way as necessary or essential. In the case of a yard, the original way this length was picked out was, I think, the distance when the arm of King Henry I of England was outstretched from the tip of his finger to his nose. If this was the length of a yard, it nevertheless will not be a necessary truth that the distance between the tip of his finger and his nose should be a yard. Maybe an accident might have happened to foreshorten his arm; that would be possible. And the reason that it's not a necessary truth is not that there might be other criteria in a 'cluster concept' of yardhood. Even a man who strictly uses King Henry's arm as his one standard of length can say, counterfactually, that if certain things had happened to the King, the exact distance between the end of one of his fingers and his nose would not have been exactly a yard. He need not be using a cluster as long as he uses the term 'yard' to pick out a certain fixed reference to be that length in all possible worlds.

These remarks show, I think, the intuitive bizarreness of a good deal of the literature on 'transworld identification' and 'counterpart theory'. For many theorists of these sorts, believing, as they do, that a 'possible world' is given to us only qualitatively, argue that Aristotle is to be 'identified in other possible worlds', or alternatively that his counterparts are to be identified, with those things in other possible worlds who most closely resemble Aristotle in his most important properties. (Lewis, for example, says: 'Your counterparts . . . resemble you . . . in important respects . . . more closely than do the other things in their worlds . . . weighted by the importance of the various respects and by the degrees of the similarities.'[32]) Some may equate the important properties with those

[32] D. Lewis, op. cit., pp. 114–15.

properties used to identify the object in the actual world.

Surely these notions are incorrect. To me Aristotle's most important properties consist in his philosophical work, and Hitler's in his murderous political role; both, as I have said, might have lacked these properties altogether. Surely there was no logical fate hanging over either Aristotle or Hitler which made it in any sense inevitable that they should have possessed the properties we regard as important to them; they could have had careers completely different from their actual ones. *Important* properties of an object need not be essential, unless 'importance' is used as a synonym for essence; and an object could have had properties very different from its most striking actual properties, or from the properties we use to identify it.

To clear up one thing which some people have asked me: When I say that a designator is rigid, and designates the same thing in all possible worlds, I mean that, as used in *our* language, it stands for that thing, when *we* talk about counterfactual situations. I don't mean, of course, that there mightn't be counterfactual situations in which in the other possible worlds people actually spoke a different language. One doesn't say that 'two plus two equals four' is contingent because people might have spoken a language in which 'two plus two equals four' meant that seven is even. Similarly, when we speak of a counterfactual situation, we speak of it in English, even if it is part of the description of that counterfactual situation that we were all speaking German in that counterfactual situation. We say, 'suppose we had all been speaking German' or 'suppose we had been using English in a nonstandard way'. Then we are describing a possible world or counterfactual situation in which people, including ourselves, did speak in a certain way different from the way we speak. But still, in describing that world, we use *English* with *our* meanings and *our* references. It is in this sense that I speak of a rigid designator as having the same

reference in all possible worlds. I also don't mean to imply that the thing designated exists in all possible worlds, just that the name refers rigidly to that thing. If you say 'suppose Hitler had never been born' then 'Hitler' refers here, still rigidly, to something that would not exist in the counterfactual situation described.

Given these remarks, this means we must cross off Thesis (6) as incorrect. The other theses have nothing to do with necessity and can survive. In particular Thesis (5) has nothing to do with necessity and it can survive. If I use the name 'Hesperus' to refer to a certain planetary body when seen in a certain celestial position in the evening, it will not therefore be a necessary truth that Hesperus is ever seen in the evening. That depends on various contingent facts about people being there to see and things like that. So even if I should say to myself that I will use 'Hesperus' to name the heavenly body I see in the evening in yonder position of the sky, it will not be necessary that Hesperus was ever seen in the evening. But it may be *a priori* in that this is how I have determined the referent. If I have determined that Hesperus is the thing that I saw in the evening over there, then I will know, just from making that determination of the referent, that if there is any Hesperus at all it's the thing I saw in the evening. This at least survives as far as the arguments we have given up to now go.

How about a theory where Thesis (6) is eliminated? Theses (2), (3), and (4) turn out to have a large class of counterinstances. Even when Theses (2)–(4) are true, Thesis (5) is usually false; the truth of Theses (3) and (4) is an empirical 'accident', which the speaker hardly knows *a priori*. That is to say, other principles really determine the speaker's reference, and the fact that the referent coincides with that determined by (2)–(4) is an 'accident', which we were in no position to know *a priori*. Only in a rare class of cases, usually initial baptisms, are all of (2)–(5) true.

What picture of naming do these Theses ((1)–(5)) give you? The picture is this. I want to name an object. I think of some way of describing it uniquely and then I go through, so to speak, a sort of mental ceremony: By 'Cicero' I shall mean the man who denounced Catiline; and that's what the reference of 'Cicero' will be. I will use 'Cicero' to designate rigidly the man who (in fact) denounced Catiline, so I can speak of possible worlds in which he did not. But still my intentions are given by first, giving some condition which uniquely determines an object, then using a certain word as a name for the object determined by this condition. Now there may be some cases in which we actually do this. Maybe, if you want to stretch and call it description, when you say: I shall call that heavenly body over there 'Hesperus'.[33] That is really a case where the theses not only are true but really even give a correct picture of how the reference is determined. Another case, if you want to call this a name, might be when the police in London use the name 'Jack' or 'Jack the Ripper' to refer to the man, whoever he is, who committed all these murders, or most of them. Then they are giving the reference of the name

[33] An even better case of determining the reference of a name by description, as opposed to ostension, is the discovery of the planet Neptune. Neptune was hypothesized as the planet which caused such and such discrepancies in the orbits of certain other planets. If Leverrier indeed gave the name 'Neptune' to the planet before it was ever seen, then he fixed the reference of 'Neptune' by means of the description just mentioned. At that time he was unable to see the planet even through a telescope. At this stage, an *a priori* material equivalence held between the statements 'Neptune exists' and 'some one planet perturbing the orbit of such and such other planets exists in such and such a position', and also such statements as 'if such and such perturbations are caused by a planet, they are caused by Neptune' had the status of *a priori* truths. Nevertheless, they were not *necessary* truths, since 'Neptune' was introduced as a name rigidly designating a certain planet. Leverrier could well have believed that if Neptune had been knocked off its course one million years earlier, it would have caused no such perturbations and even that some other object might have caused the perturbations in its place.

by a description.[34] But in many or most cases, I think the theses are false. So let's look at them.[35]

Thesis (1), as I say, is a definition. Thesis (2) says that one of the properties believed by *A* of the object, or some conjointly, are believed to pick out some individual uniquely. A sort of example people have in mind is just what I said: I shall use the term 'Cicero' to denote the man who denounced Catiline (or first denounced him in public, to make it unique). This picks out an object uniquely in this particular reference. Even some writers such as Ziff in *Semantic Analysis*, who don't believe that names have meaning in any sense, think that this is a good picture of the way reference can be determined.

Let's see if Thesis (2) is true. It seems, in some *a priori* way, that it's got to be true, because if you don't think that the properties you have in mind pick out anyone uniquely—let's say they're all satisfied by two people—then how can you say which one of them you're talking about? There seem to be no grounds for saying you're talking about the one rather than about the other. Usually the properties in question are supposed to be some famous deeds of the person in question. For example, Cicero was the man who denounced Catiline. The average person, according to this, when he refers to Cicero, is

[34] Following Donnellan's remarks on definite descriptions, we should add that in some cases, an object may be identified, and the reference of a name fixed, using a description which may turn out to be false of its object. The case where the reference of 'Phosphorus' is determined as the 'morning star', which later turns out not to be a star, is an obvious example. In such cases, the description which fixes the reference clearly is in no sense known *a priori* to hold of the object, though a more cautious substitute may be. If such a more cautious substitute is available, it is really the substitute which fixes the reference in the sense intended in the text.

[35] Some of the theses are sloppily stated in respect of fussy matters like use of quotation marks and related details. (For example, Theses (5) and (6), as stated, presuppose that the speaker's language is English.) Since the purport of the theses is clear, and they are false anyway, I have not bothered to set these things straight.

saying something like 'the man who denounced Catiline' and thus has picked out a certain man uniquely. It is a tribute to the education of philosophers that they have held this thesis for such a long time. In fact, most people, when they think of Cicero, just think of *a famous Roman orator*, without any pretension to think either that there was only one famous Roman orator or that one must know something else about Cicero to have a referent for the name. Consider Richard Feynman, to whom many of us are able to refer. He is a leading contemporary theoretical physicist. Everyone *here* (I'm sure!) can state the contents of one of Feynman's theories so as to differentiate him from Gell-Mann. However, the man in the street, not possessing these abilities, may still use the name 'Feynman'. When asked he will say: well he's a physicist or something. He may not think that this picks out anyone uniquely. I still think he uses the name 'Feynman' as a name for Feynman.

But let's look at some of the cases where we do have a description to pick out someone uniquely. Let's say, for example, that we know that Cicero was the man who first denounced Catiline. Well, that's good. That really picks someone out uniquely. However, there is a problem, because this description contains another name, namely 'Catiline'. We must be sure that we satisfy the conditions in such a way as to avoid violating the noncircularity condition here. In particular, we must not say that Catiline was the man denounced by Cicero. If we do this, we will really not be picking out anything uniquely, we will simply be picking out a pair of objects *A* and *B*, such that *A* denounced *B*. We do not think that this was the only pair where such denunciations ever occurred; so we had better add some other conditions in order to satisfy the uniqueness condition.

If we say Einstein was the man who discovered the theory of relativity, that certainly picks out someone uniquely. One can

be sure, as I said, that everyone *here* can make a compact and independent statement of this theory and so pick out Einstein uniquely; but many people actually don't know enough about this stuff, so when asked what the theory of relativity is, they will say: 'Einstein's theory', and thus be led into the most straightforward sort of vicious circle.

So Thesis (2), in a straightforward way, fails to be satisfied when we say Feynman is a famous physicist without attributing anything else to Feynman. In another way it may not be satisfied in the proper way even when it is satisfied: If we say Einstein was 'the man who discovered relativity theory', that does pick someone out uniquely; but it may not pick him out in such a way as to satisfy the noncircularity condition, because the theory of relativity may in turn be picked out as 'Einstein's theory'. So Thesis (2) seems to be false.

By changing the conditions φ from those usually associated with names by philosophers, one could try to improve the theory. There have been various ways I've heard; maybe I'll discuss these later on. Usually they think of famous achievements of the man named. Certainly in the case of famous achievements, the theory doesn't work. Some student of mine once said, 'Well, Einstein discovered the theory of relativity'; and he determined the reference of 'the theory of relativity' independently by referring to an encyclopedia which would give the details of the theory. (This is what is called a transcendental deduction of the existence of encyclopedias.) But it seems to me that, even if someone has heard of encyclopedias, it really is not essential for his reference that he should know whether this theory is given in detail in any encyclopedia. The reference might work even if there had been no encyclopedias at all.

Let's go on to Thesis (3): If most of the φ's, suitably weighted, are satisfied by a unique object y, then y is the referent of the name for the speaker. Now, since we have already established

that Thesis (2) is wrong, why should any of the rest work? The whole theory depended on always being able to specify unique conditions which are satisfied. But still we can look at the other theses. The picture associated with the theory is that only by giving some unique properties can you know who someone is and thus know what the reference of your name is. Well, I won't go into the question of knowing who someone is. It's really very puzzling. I think you *do* know who Cicero is if you just can answer that he's a famous Roman orator. Strangely enough, if you know that Einstein discovered the theory of relativity and nothing about that theory, you can both know who Einstein is, namely the discoverer of the theory of relativity, and who discovered the theory of relativity, namely Einstein, on the basis of this knowledge. This seems to be a blatant violation of some sort of noncircularity condition; but it is the way we talk. It therefore would seem that a picture which suggests this condition must be the wrong picture.

Suppose most of the φ's are in fact satisfied by a unique object. Is that object necessarily the referent of 'X' for A? Let's suppose someone says that Gödel is the man who proved the incompleteness of arithmetic, and this man is suitably well educated and is even able to give an independent account of the incompleteness theorem. He doesn't just say, 'Well, that's Gödel's theorem', or whatever. He actually states a certain theorem, which he attributes to Gödel as the discoverer. Is it the case, then, that if most of the φ's are satisfied by a unique object y, then y is the referent of the name 'X' for A? Let's take a simple case. In the case of Gödel that's practically the only thing many people have heard about him—that he discovered the incompleteness of arithmetic. Does it follow that whoever discovered the incompleteness of arithmetic is the referent of 'Gödel'?

Imagine the following blatantly fictional situation. (I hope Professor Gödel is not present.) Suppose that Gödel was not in

fact the author of this theorem. A man named 'Schmidt', whose body was found in Vienna under mysterious circumstances many years ago, actually did the work in question. His friend Gödel somehow got hold of the manuscript and it was thereafter attributed to Gödel. On the view in question, then, when our ordinary man uses the name 'Gödel', he really means to refer to Schmidt, because Schmidt is the unique person satisfying the description, 'the man who discovered the incompleteness of arithmetic'. Of course you might try changing it to 'the man who *published* the discovery of the incompleteness of arithmetic'. By changing the story a little further one can make even this formulation false. Anyway, most people might not even know whether the thing was published or got around by word of mouth. Let's stick to 'the man who discovered the incompleteness of arithmetic'. So, since the man who discovered the incompleteness of arithmetic is in fact Schmidt, we, when we talk about 'Gödel', are in fact always referring to Schmidt. But it seems to me that we are not. We simply are not. One reply, which I will discuss later, might be: You should say instead, 'the man to whom the incompleteness of arithmetic is commonly attributed', or something like that. Let's see what we can do with that later.

But it may seem to many of you that this is a very odd example, or that such a situation occurs rarely. This also is a tribute to the education of philosophers. Very often we use a name on the basis of considerable misinformation. The case of mathematics used in the fictive example is a good case in point. What do we know about Peano? What many people in this room may 'know' about Peano is that he was the discoverer of certain axioms which characterize the sequence of natural numbers, the so-called 'Peano axioms'. Probably some people can even state them. I have been told that these axioms were not first discovered by Peano but by Dedekind. Peano was of course not a dishonest man. I am told that his footnotes

include a credit to Dedekind. Somehow the footnote has been ignored. So on the theory in question the term 'Peano', as we use it, really refers to—now that you've heard it you see that you were really all the time talking about—Dedekind. But you were not. Such illustrations could be multiplied indefinitely.

Even worse misconceptions, of course, occur to the layman. In a previous example I supposed people to identify Einstein by reference to his work on relativity. Actually, I often used to hear that Einstein's most famous achievement was the invention of the atomic bomb. So when we refer to Einstein, we refer to the inventor of the atomic bomb. But this is not so. Columbus was the first man to realize that the earth was round. He was also the first European to land in the western hemisphere. Probably none of these things are true, and therefore, when people use the term 'Columbus' they really refer to some Greek if they use the roundness of the earth, or to some Norseman, perhaps, if they use the 'discovery of America'. But they don't. So it does not seem that if most of the φ's are satisfied by a unique object y, then y is the referent of the name. This seems simply to be false.[36]

[36] The cluster-of-descriptions theory of naming would make 'Peano discovered the axioms for number theory' express a trivial truth, not a misconception, and similarly for other misconceptions about the history of science. Some who have conceded such cases to me have argued that there are *other* uses of the same proper names satisfying the cluster theory. For example, it is argued, if we say, 'Gödel proved the incompleteness of arithmetic,' we are, of course, referring to Gödel, not to Schmidt. But, if we say, 'Gödel relied on a diagonal argument in this step of the proof,' don't we here, perhaps, refer to *whoever proved the theorem*? Similarly, if someone asks, 'What did Aristotle (or Shakespeare) have in mind here?', isn't he talking about the author of the passage in question, whoever he is? By analogy to Donnellan's usage for descriptions, this might be called an "attributive' use of proper names. If this is so, then assuming the Gödel-Schmidt story, the sentence 'Gödel proved the incompleteness theorem' is false, but 'Gödel used a diagonal argument in the proof' is (at least in some contexts) true, and the reference of the name 'Gödel' is ambiguous. Since some counterexamples remain, the

Thesis (4): If the vote yields no unique object the name does not refer. Really this case has been covered before—has been covered in my previous examples. First, the vote may not yield a *unique* object, as in the case of Cicero or Feynman. Secondly, suppose it yields *no* object, that nothing satisfies most, or even any, substantial number, of the φ's. Does that mean the name doesn't refer? No: in the same way that you may have false beliefs about a person which may actually be true of someone else, so you may have false beliefs which are true of absolutely no one. And these may constitute the totality of your beliefs. Suppose, to vary the example about Gödel, no one had discovered the incompleteness of arithmetic—perhaps the proof simply materialized by a random scattering of atoms on a piece of paper—the man Gödel being lucky enough to have been present when this improbable event occurred. Further, suppose arithmetic is in fact complete. One wouldn't really expect a random scattering of atoms to produce a correct proof. A subtle error, unknown through the decades, has still been unnoticed—or perhaps not actually unnoticed, but the friends of Gödel. . . . So even if the conditions are not satisfied

cluster-of-descriptions theory would still, in general, be false, which was my main point in the text; but it would be applicable in a wider class of cases than I thought. I think, however, that no such ambiguity need be postulated. It is, perhaps, true that sometimes when someone uses the name 'Gödel', his main interest is in whoever proved the theorem, and *perhaps*, in some sense, he 'refers' to him. I do not think that this case is different from the case of Smith and Jones in n. 3, p. 25. If I mistake Jones for Smith, I may *refer* (in an appropriate sense) to Jones when I say that Smith is raking the leaves; nevertheless I do not use 'Smith' ambiguously, as a name sometimes of Smith and sometimes of Jones, but univocally as a name of Smith. Similarly, if I erroneously think that Aristotle wrote such-and-such passage, I may perhaps sometimes use 'Aristotle' to *refer* to the actual author of the passage, even though there is no ambiguity in my use of the name. In both cases, I will withdraw my original statement, and my original use of the name, if apprised of the facts. Recall that, in these lectures, 'referent' is used in the technical sense of the thing named by a name (or uniquely satisfying a description), and there should be no confusion.

by a unique object the name may still refer. I gave you the case of Jonah last week. Biblical scholars, as I said, think that Jonah really existed. It isn't because they think that someone ever was swallowed by a big fish or even went to Nineveh to preach. These conditions may be true of no one whatsoever and yet the name 'Jonah' really has a referent. In the case above of Einstein's invention of the bomb, possibly no one really deserves to be called the 'inventor' of the device.

Thesis 5 says that the statement 'If X exists, then X has most of the φ's', is *a priori* true for A. Notice that even in a case where (3) and (4) *happen* to be true, a typical speaker hardly knows *a priori* that they are, as required by the theory. I *think* that my belief about Gödel *is* in fact correct and that the 'Schmidt' story is just a fantasy. But the belief hardly constitutes *a priori* knowledge.

What's going on here? Can we rescue the theory?[37] First, one may try and vary these descriptions—not think of the famous achievements of a man but, let's say, of something else, and try and use that as our description. Maybe by enough futzing around someone might eventually get something out

[37] It has been suggested to me that someone might argue that a name is associated with a 'referential' use of a description in Donnellan's sense. For example, although we identify Gödel as the author of the incompleteness theorem, we are talking about him even if he turns out not to have proved the theorem. Theses (2)–(6) could then fail; but nevertheless each name would abbreviate a description, though the role of description in naming would differ radically from that imagined by Frege and Russell. As I have said above, I am inclined to reject Donnellan's formulation of the notion of referential definite description. Even if Donnellan's analysis is accepted, however, it is clear that the present proposal should not be. For a referential definite description, such as 'the man drinking champagne', is typically withdrawn when the speaker realizes that it does not apply to its object. If a Gödelian fraud were exposed, Gödel would no longer be called 'the author of the incompleteness theorem' but he would still be called 'Gödel'. The name, therefore, does not abbreviate the description.

of this;[38] however, most of the attempts that one tries are open to counterexamples or other objections. Let me give an example of this. In the case of Gödel one may say, 'Well, "Gödel" doesn't mean "the man who proved the incompleteness of arithmetic" '. Look, all we really know is that most people *think* that Gödel proved the incompleteness of arithmetic, that Gödel is the man to whom the incompleteness of arithmetic is commonly attributed. So when I determine the referent of the name 'Gödel', I don't say to myself, 'by "Gödel" I shall mean "the man who proved the incompleteness of arithmetic, whoever he is" '. That might turn out to be Schmidt or Post. But instead I shall mean 'the man who most people *think* proved the incompleteness of arithmetic'.

Is this right? First, it seems to me that it's open to counterexamples of the same type as I gave before, though the counterexamples may be more recherché. Suppose, in the case of Peano mentioned previously, unbeknownst to the speaker, most people (at least by now) thoroughly realize that the number-theoretic axioms should not be attributed to him. Most people don't credit them to Peano but now correctly ascribe them to Dedekind. So then even the man to whom this thing is commonly attributed will still be Dedekind and not Peano. Still, the speaker, having picked up the old outmoded

[38] As Robert Nozick pointed out to me, there is a sense in which a description theory must be trivially true if any theory of the reference of names, spelled out in terms independent of the notion of reference, is available. For if such a theory gives conditions under which an object is to be the referent of a name, then it of course uniquely satisfies these conditions. Since I am not pretending to give any theory which eliminates the notion of reference in this sense, I am not aware of any such trivial fulfillment of the description theory and doubt that one exists. (A description using the notion of the reference of a name is easily available but circular, as we saw in our discussion of Kneale.) If any such trivial fulfillment were available, however, the arguments I have given show that the description must be one of a completely different sort from that supposed by Frege, Russell, Searle, Strawson and other advocates of the description theory.

belief, may still be referring to Peano, and hold a false belief about Peano, not a true belief about Dedekind.

But second, and perhaps more significantly, such a criterion violates the noncircularity condition. How is this? It is true that most of us think that Gödel proved the incompleteness of arithmetic. Why is this so? We certainly say, and sincerely, 'Gödel proved the incompleteness of arithmetic'. Does it follow from that that we believe that Gödel proved the incompleteness of arithmetic—that we attribute the incompleteness of arithmetic to this man? No. Not just from that. We have to be *referring to Gödel* when we say 'Gödel proved the incompleteness of arithmetic'. If, in fact, we were always referring to Schmidt, then we would be attributing the incompleteness of arithmetic to Schmidt and not to Gödel—if we used the sound 'Gödel' as the name of the man whom I am calling 'Schmidt'.

But we do in fact refer to Gödel. How do we do this? Well, not by saying to ourselves, 'By "Gödel" I shall mean the man to whom the incompleteness of arithmetic is commonly attributed'. If we did that we would run into a circle. Here we are all in this room. Actually in this institution[39] some people have met the man, but in many institutions this is not so. All of us in the community are trying to determine the reference by saying 'Gödel is to be the man to whom the incompleteness of arithmetic is commonly attributed'. None of us will get started with any attribution unless there is some independent criterion for the reference of the name other than 'the man to whom the incompleteness of arithmetic is commonly attributed'. Otherwise all we will be saying is, 'We attribute this achievement to the man to whom we attribute it', without saying who that man is, without giving any independent criterion of the reference, and so the determination will be circular. This then is a violation of the condition I have

[39] Princeton University.

marked '*C*', and cannot be used in any theory of reference.

Of course you might try to avoid circularity by passing the buck. This is mentioned by Strawson, who says in his footnote on these matters that one man's reference may derive from another's.

> The identifying description, though it must not include a reference to the speaker's own reference to the particular in question, may include a reference to another's reference to that particular. If a putatively identifying description is of this latter kind, then, indeed, the question, whether it is a genuinely identifying description, turns on the question, whether the reference it refers to is itself a genuinely identifying reference. So one reference may borrow its credentials, as a genuinely identifying reference, from another; and that from another. But this regress is not infinite.[40]

I may then say, 'Look, by "Gödel" I shall mean the man Joe thinks proved the incompleteness of arithmetic'. Joe may then pass the thing over to Harry. One has to be very careful that this doesn't come round in a circle. Is one really sure that this won't happen? If you could be sure yourself of knowing such a chain, and that everyone else in the chain is using the proper conditions and so is not getting out of it, then maybe you could get back to the man by referring to such a chain in that way, borrowing the references one by one. However, although in general such chains do exist for a living man, you won't know what the chain is. You won't be sure what descriptions the other man is using, so the thing won't go into a circle, or whether by appealing to Joe you won't get back to the right man at all. So you cannot use this as your identifying description with any confidence. You may not even remember from whom you heard of Gödel.

What is the true picture of what's going on? Maybe reference doesn't really take place at all! After all, we don't really know

[40] Strawson, op. cit., p. 182 n.

that any of the properties we use to identify the man are right. We don't know that they pick out a unique object. So what *does* make my use of 'Cicero' into a name of *him*? The picture which leads to the cluster-of-descriptions theory is something like this: One is isolated in a room; the entire community of other speakers, everything else, could disappear; and one determines the reference for himself by saying—'By "Gödel" I shall mean the man, whoever he is, who proved the incompleteness of arithmetic'. Now you can do this if you want to. There's nothing really preventing it. You can just stick to that determination. If that's what you do, then if Schmidt discovered the incompleteness of arithmetic you *do* refer to him when you say 'Gödel did such and such'.

But that's not what most of us do. Someone, let's say, a baby, is born; his parents call him by a certain name. They talk about him to their friends. Other people meet him. Through various sorts of talk the name is spread from link to link as if by a chain. A speaker who is on the far end of this chain, who has heard about, say Richard Feynman, in the market place or elsewhere, may be referring to Richard Feynman even though he can't remember from whom he first heard of Feynman or from whom he ever heard of Feynman. He knows that Feynman is a famous physicist. A certain passage of communication reaching ultimately to the man himself does reach the speaker. He then is referring to Feynman even though he can't identify him uniquely. He doesn't know what a Feynman diagram is, he doesn't know what the Feynman theory of pair production and annihilation is. Not only that: he'd have trouble distinguishing between Gell-Mann and Feynman. So he doesn't have to know these things, but, instead, a chain of communication going back to Feynman himself has been established, by virtue of his membership in a community which passed the name on from link to link, not by a ceremony that he makes in private in his study: 'By "Feynman" I shall

mean the man who did such and such and such and such'.

How does this view differ from Strawson's suggestion, mentioned before, that one identifying reference may borrow its credentials from another? Certainly Strawson had a good insight in the passage quoted; on the other hand, he certainly shows a difference at least in emphasis from the picture I advocate, since he confines the remark to a footnote. The main text advocates the cluster-of-descriptions theory. Just because Strawson makes his remark in the context of a description theory, his view therefore differs from mine in one important respect. Strawson apparently requires that the speaker must *know* from whom he got his reference, so that he can say: 'By "Gödel" I mean the man *Jones* calls "Gödel" '. If he does not remember how he picked up the reference, he cannot give such a description. The present theory sets no such requirement. As I said, I may well not remember from whom I heard of Gödel, and I may think I remember from which people I heard the name, but wrongly.

These considerations show that the view advocated here can lead to consequences which actually *diverge* from those of Strawson's footnote. Suppose that the speaker has heard the name 'Cicero' from Smith and others, who use the name to refer to a famous Roman orator. He later thinks, however, that he picked up the name from Jones, who (unknown to the speaker) uses 'Cicero' as the name of a notorious German spy and has never heard of any orators of the ancient world. Then, according to Strawson's paradigm, the speaker must determine his reference by the resolution, 'I shall use "Cicero" to refer to the man whom Jones calls by that name', while on the present view, the referent will be the orator in spite of the speaker's false impression about where he picked up the name. The point is that Strawson, trying to fit the chain of communication view into the description theory, relies on what the speaker *thinks* was the source of his reference. If the speaker has forgotten his source, the description Strawson uses is unavailable to him; if he misremembers it, Strawson's paradigm can give the wrong results. On our view, it is not how the speaker thinks he got the reference, but the actual chain of communication, which is relevant.

I think I said the other time that philosophical theories are in danger of being false, and so I wasn't going to present an alternative theory. Have I just done so? Well, in a way; but my characterization has been far less specific than a real set of necessary and sufficient conditions for reference would be. Obviously the name is passed on from link to link. But of course not every sort of causal chain reaching from me to a certain man will do for me to make a reference. There may be a causal chain from our use of the term 'Santa Claus' to a certain historical saint, but still the children, when they use this, by this time probably do not refer to that saint. So other conditions must be satisfied in order to make this into a really rigorous theory of reference. I don't know that I'm going to do this because, first, I'm sort of too lazy at the moment; secondly, rather than giving a set of necessary and sufficient conditions which will work for a term like reference, I want to present just a *better picture* than the picture presented by the received views.

Haven't I been very unfair to the description theory? Here I have stated it very precisely—more precisely, perhaps, than it has been stated by any of its advocates. So then it's easy to refute. Maybe if I tried to state mine with sufficient precision in the form of six or seven or eight theses, it would also turn out that when you examine the theses one by one, they will all be false. That might even be so, but the difference is this. What I think the examples I've given show is not simply that there's some technical error here or some mistake there, but that the whole picture given by this theory of how reference is determined seems to be wrong from the fundamentals. It

seems to be wrong to think that we give ourselves some properties which somehow qualitatively uniquely pick out an object and determine our reference in that manner. What I am trying to present is a better picture—a picture which, if more details were to be filled in, might be refined so as to give more exact conditions for reference to take place.

One might never reach a set of necessary and sufficient conditions. I don't know, I'm always sympathetic to Bishop Butler's 'Everything is what it is and not another thing'—in the nontrivial sense that philosophical analyses of some concept like reference, in completely different terms which make no mention of reference, are very apt to fail. Of course in any particular case when one is given an analysis one has to look at it and see whether it is true or false. One can't just cite this maxim to oneself and then turn the page. But more cautiously, I want to present a better picture without giving a set of necessary and sufficient conditions for reference. Such conditions would be very complicated, but what is true is that it's in virtue of our connection with other speakers in the community, going back to the referent himself, that we refer to a certain man.

There may be some cases where the description picture is true, where some man really gives a name by going into the privacy of his room and saying that the referent is to be the unique thing with certain identifying properties. 'Jack the Ripper' was a possible example which I gave. Another was 'Hesperus'. Yet another case which can be forced into this description is that of meeting someone and being told his name. Except for a belief in the description theory, in its importance in other cases, one probably wouldn't think that that was a case of giving oneself a description, i.e., 'the guy I'm just meeting now'. But one can put it in these terms if one wishes, and if one has never heard the name in any other way. Of course, if you're introduced to a man and told, 'That's

Einstein', you've heard of him before, it may be wrong, and so on. But maybe in some cases such a paradigm works—especially for the man who first gives someone or something a name. Or he points to a star and says, 'That is to be Alpha Centauri'. So he can really make himself this ceremony: 'By "Alpha Centauri" I shall mean the star right over there with such and such coordinates'. But in general this picture fails. In general our reference depends not just on what we think ourselves, but on other people in the community, the history of how the name reached one, and things like that. It is by following such a history that one gets to the reference.

More exact conditions are very complicated to give. They seem in a way somehow different in the case of a famous man and one who isn't so famous. For example, a teacher tells his class that Newton was famous for being the first man to think there's a force pulling things to the earth; I think that's what little kids think Newton's greatest achievement was. I won't say what the merits of such an achievement would be, but, anyway, we may suppose that just being told that this was the sole content of Newton's discovery gives the students a false belief *about Newton*, even though they have never heard of him before. If, on the other hand,[41] the teacher uses the name 'George Smith'—a man by that name is actually his next door neighbor—and says that George Smith first squared the circle, does it follow from this that the students have a false belief about the teacher's neighbor? The teacher doesn't tell them that Smith is his neighbor, nor does he believe Smith first squared the circle. He isn't particularly trying to get any belief *about the neighbor* into the students' heads. He tries to inculcate the belief that there was a man who squared the circle, but not a belief about any particular man—he just pulls out the first name that occurs to him—as it happens, he uses his neighbor's name. It doesn't seem clear in that case that the

[41] The essential points of this example were suggested by Richard Miller.

students have a false belief about the neighbor, even though there is a causal chain going back to the neighbor. I am not sure about this. At any rate more refinements need to be added to make this even begin to be a set of necessary and sufficient conditions. In that sense it's not a theory, but is supposed to give a better picture of what is actually going on.

A rough statement of a theory might be the following: An initial 'baptism' takes place. Here the object may be named by ostension, or the reference of the name may be fixed by a description.[42] When the name is 'passed from link to link', the receiver of the name must, I think, intend when he learns it to use it with the same reference as the man from whom he heard it. If I hear the name 'Napoleon' and decide it would be a nice name for my pet aardvark, I do not satisfy this condition.[43] (Perhaps it is some such failure to keep the reference

[42] A good example of a baptism whose reference was fixed by means of a description was that of naming Neptune in n. 33, p. 79. The case of a baptism by ostension can perhaps be subsumed under the description concept also. Thus the primary applicability of the description theory is to cases of initial baptism. Descriptions are also used to fix a reference in cases of designation which are similar to naming except that the terms introduced are not usually called 'names'. The terms 'one meter', '100 degrees Centigrade', have already been given as examples, and other examples will be given later in these lectures. Two things should be emphasized concerning the case of introducing a name via a description in an initial baptism. First, the description used is not synonymous with the name it introduces but rather fixes its reference. Here we differ from the usual description theorists. Second, most cases of initial baptism are far from those which originally inspired the description theory. Usually a baptizer is acquainted in some sense with the object he names and is able to name it ostensively. Now the inspiration of the description theory lay in the fact that we can often use names of famous figures of the past who are long dead and with whom no living person is acquainted; and it is precisely these cases which, on our view, cannot be correctly explained by a description theory.

[43] I can transmit the name of the aardvark to other people. For each of these people, as for me, there will be a certain sort of causal or historical connection between my use of the name and the Emperor of the French, but not one of the required type.

fixed which accounts for the divergence of present uses of 'Santa Claus' from the alleged original use.)

Notice that the preceding outline hardly *eliminates* the notion of reference; on the contrary, it takes the notion of intending to use the same reference as a given. There is also an appeal to an initial baptism which is explained in terms either of fixing a reference by a description, or ostension (if ostension is not to be subsumed under the other category).[44] (Perhaps there are other possibilities for initial baptisms.) Further, the George Smith case casts some doubt as to the sufficiency of the conditions. Even if the teacher does refer to his neighbor, is it clear that he has passed on his reference to the pupils? Why shouldn't their belief be about any other man named 'George Smith'? If he says that Newton was hit by an apple, somehow his task of transmitting a reference is easier, since he has communicated a common misconception about Newton.

To repeat, I may not have presented a theory, but I do think that I have presented a better picture than that given by description theorists.

I think the next topic I shall want to talk about is that of statements of identity. Are these necessary or contingent? The matter has been in some dispute in recent philosophy. First,

[44] Once we realize that the description used to fix the reference of a name is not synonymous with it, then the description theory can be regarded as presupposing the notion of naming or reference. The requirement I made that the description used not itself involve the notion of reference in a circular way is something else and is crucial if the description theory is to have any value at all. The reason is that the description theorist supposes that each speaker essentially uses the description he gives in an initial act of naming to determine his reference. Clearly, if he introduces the name 'Cicero' by the determination, 'By "Cicero" I shall refer to the man I call "Cicero",' he has by this ceremony determined no reference at all.

Not all description theorists thought that they were eliminating the notion of reference altogether. Perhaps some realized that some notion of ostension, or primitive reference, is required to back it up. Certainly Russell did.

everyone agrees that descriptions can be used to make contingent identity statements. If it is true that the man who invented bifocals was the first Postmaster General of the United States—that these were one and the same—it's contingently true. That is, it might have been the case that one man invented bifocals and another was the first Postmaster General of the United States. So certainly when you make identity statements using descriptions—when you say 'the x such that φx and the x such that ψx are one and the same'—that can be a contingent fact. But philosophers have been interested also in the question of identity statements between names. When we say 'Hesperus is Phosphorus' or 'Cicero is Tully', is what we are saying necessary or contingent? Further, they've been interested in another type of identity statement, which comes from scientific theory. We identify, for example, light with electromagnetic radiation between certain limits of wavelengths, or with a stream of photons. We identify heat with the motion of molecules; sound with a certain sort of wave disturbance in the air; and so on. Concerning such statements the following thesis is commonly held. First, that these are obviously contingent identities: we've found out that light is a stream of photons, but of course it might not have been a stream of photons. Heat is in fact the motion of molecules; we found that out, but heat might not have been the motion of molecules. Secondly, many philosophers feel damned lucky that these examples are around. Now, why? These philosophers, whose views are expounded in a vast literature, hold to a thesis called 'the identity thesis' with respect to some psychological concepts. They think, say, that pain is just a certain material state of the brain or of the body, or what have you—say the stimulation of C-fibers. (It doesn't matter what.) Some people have then objected, 'Well, look, there's perhaps a *correlation* between pain and these states of the body; but this must just be a contingent correlation between two different things, because

it was an empirical discovery that this correlation ever held. Therefore, by "pain" we must mean something different from this state of the body or brain; and, therefore, they must be two different things.'

Then it's said, 'Ah, but you see, this is wrong! Everyone knows that there can be contingent identities.' First, as in the bifocals and Postmaster General case, which I have mentioned before. Second, in the case, believed closer to the present paradigm, of theoretical identifications, such as light and a stream of photons, or water and a certain compound of hydrogen and oxygen. These are all contingent identities. They might have been false. It's no surprise, therefore, that it can be true as a matter of contingent fact and not of any necessity that feeling pain, or seeing red, is just a certain state of the human body. Such psychophysical identifications can be contingent facts just as the other identities are contingent facts. And of course there are widespread motivations—ideological, or just not wanting to have the 'nomological dangler' of mysterious connections not accounted for by the laws of physics, one to one correlations between two different kinds of thing, material states, and things of an entirely different kind, which lead people to want to believe this thesis.

I guess the main thing I'll talk about first is identity statements between names. But I hold the following about the general case. First, that characteristic theoretical identifications like 'Heat is the motion of molecules', are not contingent truths but necessary truths, and here of course I don't mean just physically necessary, but necessary in the highest degree—whatever that means. (Physical necessity, *might* turn out to be necessity in the highest degree. But that's a question which I don't wish to prejudge. At least for this sort of example, it might be that when something's physically necessary, it always is necessary *tout court*.) Second, that the way in which these have turned out to be necessary truths does not seem to me to

be a way in which the mind-brain identities could turn out to be either necessary or contingently true. So this analogy has to go. It's hard to see what to put in its place. It's hard to see therefore how to avoid concluding that the two are actually different.

Let me go back to the more mundane case about proper names. This is already mysterious enough. There's a dispute about this between Quine and Ruth Barcan Marcus.[45] Marcus says that identities between names are necessary. If someone thinks that Cicero is Tully, and really uses 'Cicero' and 'Tully' as names, he is thereby committed to holding that his belief is a necessary truth. She uses the term 'mere tag'. Quine replies as follows, 'We may tag the planet Venus, some fine evening, with the proper name "Hesperus". We may tag the same planet again, some day before sunrise, with the proper name "Phosphorus". When we discover that we have tagged the same planet twice our discovery is empirical. And not because the proper names were descriptions.'[46] First, as Quine says when we discovered that we tagged the same planet twice, our discovery was empirical. Another example I think Quine gives in another book is that the same mountain seen from Nepal and from Tibet, or something like that, is from one angle called 'Mt. Everest' (you've heard of that); from another it's supposed to be called 'Gaurisanker'. It can actually be an empirical discovery that Gaurisanker is Everest. (Quine says that the example is actually false. He got the example from Erwin Schrödinger. You wouldn't think the inventor of wave mechanics got things that wrong. I don't know where the mistake is supposed to come from. One could certainly imagine this situation as having been the case; and it's another

[45] Ruth Barcan Marcus, 'Modalities and Intensional Languages' (comments by W. V. Quine, plus discussion) *Boston Studies in the Philosophy of Science*, volume I, Reidel, Dordrecht, Holland, 1963, pp. 77–116.

[46] p. 101.

good illustration of the sort of thing that Quine has in mind.)

What about it? I wanted to find a good quote on the other side from Marcus in this book but I am having trouble locating one. Being present at that discussion, I remember[47] that she advocated the view that if you really have names, a good dictionary should be able to tell you whether they have the same reference. So someone should be able, by looking in the dictionary, to say that Hesperus and Phosphorus are the same. Now this does not seem to be true. It does seem, to many people, to be a consequence of the view that identities between names are necessary. Therefore the view that identity statements between names are necessary has usually been rejected. Russell's conclusion was somewhat different. He did think there should never be any empirical question whether two names have the same reference. This isn't satisfied for ordinary names, but it is satisfied when you're naming your own sense datum, or something like that. You say, 'Here, this, and that (designating the same sense datum by both demonstratives).' So you can tell without empirical investigation that you're naming the same thing twice; the conditions are satisfied. Since this won't apply to ordinary cases of naming, ordinary 'names' cannot be genuine names.

What should we think about this? First, it's true that someone can use the name 'Cicero' to refer to Cicero and the name 'Tully' to refer to Cicero also, and not know that Cicero is Tully. So it seems that we do not necessarily know *a priori* that an identity statement between names is true. It doesn't follow from this that the statement so expressed is a contingent one if true. This is what I've emphasized in my first lecture. There is a very strong feeling that leads one to think that, if you can't know something by *a priori* ratiocination, then it's got to be contingent: it might have turned out otherwise; but nevertheless I think this feeling is wrong.

[47] p. 115.

Bibliothek

Let's suppose we refer to the same heavenly body twice, as 'Hesperus' and 'Phosphorus'. We say: Hesperus is that star over there in the evening; Phosphorus is that star over there in the morning. Actually, Hesperus is Phosphorus. Are there really circumstances under which Hesperus wouldn't have been Phosphorus? Supposing that Hesperus is Phosphorus, let's try to describe a possible situation in which it would not have been. Well, it's easy. Someone goes by and he calls two *different* stars 'Hesperus' and 'Phosphorus'. It may even be under the same conditions as prevailed when we introduced the names 'Hesperus' and 'Phosphorus'. But are those circumstances in which Hesperus is not Phosphorus or would not have been Phosphorus? It seems to me that they are not.

Now, of course I'm committed to saying that they're not, by saying that such terms as 'Hesperus' and 'Phosphorus', when used as names, are rigid designators. They refer in every possible world to the planet Venus. Therefore, in that possible world too, the planet Venus is the planet Venus and it doesn't matter what any other person has said in this other possible world. How should *we* describe this situation? He can't have pointed to Venus twice, and in the one case called it 'Hesperus' and in the other 'Phosphorus', as we did. If he did so, then 'Hesperus is Phosphorus' would have been true in that situation too. He pointed maybe neither time to the planet Venus—at least one time he didn't point to the planet Venus, let's say when he pointed to the body he called 'Phosphorus'. Then in that case we can certainly say that the name 'Phosphorus' might not have referred to Phosphorus. We can even say that in the very position when viewed in the morning that we found Phosphorus, it might have been the case that Phosphorus was not there—that something else was there, and that even, under certain circumstances it would have been *called* 'Phosphorus'. But that still is not a case in which Phosphorus was not Hesperus. There might be a possible world in

which, a possible counterfactual situation in which, 'Hesperus' and 'Phosphorus' weren't names of the things they in fact are names of. Someone, if he did determine their reference by identifying descriptions, might even have used the very identifying descriptions we used. But still that's not a case in which Hesperus wasn't Phosphorus. For there couldn't have been such a case, given that Hesperus is Phosphorus.

Now this seems very strange because in advance, we are inclined to say, the answer to the question whether Hesperus is Phosphorus might have turned out either way. So aren't there really two possible worlds—one in which Hesperus was Phosphorus, the other in which Hesperus wasn't Phosphorus—in advance of our discovering that these were the same? First, there's one sense in which things might turn out either way, in which it's clear that that doesn't imply that the way it finally turns out isn't necessary. For example, the four color theorem might turn out to be true and might turn out to be false. It might turn out either way. It still doesn't mean that the way it turns out is not necessary. Obviously, the 'might' here is purely 'epistemic'—it merely expresses our present state of ignorance, or uncertainty.

But it seems that in the Hesperus-Phosphorus case, something even stronger is true. The evidence I have before I know that Hesperus is Phosphorus is that I see a certain star or a certain heavenly body in the evening and call it 'Hesperus', and in the morning and call it 'Phosphorus'. I know these things. There certainly is a possible world in which a man should have seen a certain star at a certain position in the evening and called it 'Hesperus' and a certain star in the morning and called it 'Phosphorus'; and should have concluded—should have found out by empirical investigation—that he names two different stars, or two different heavenly bodies. At least one of these stars or heavenly bodies was not Phosphorus, otherwise it couldn't have come out that way. But that's true. And so it's

true that given the evidence that someone has antecedent to his empirical investigation, he can be placed in a sense in exactly the same situation, that is a qualitatively identical epistemic situation, and call two heavenly bodies 'Hesperus' and 'Phosphorus', without their being identical. So in that sense we can say that it might have turned out either way. Not that it might have turned out either way as to Hesperus's being Phosphorus. Though for all we knew in advance, Hesperus wasn't Phosphorus, that couldn't have turned out any other way, in a sense. But being put in a situation where we have exactly the same evidence, qualitatively speaking, it could have turned out that Hesperus was not Phosphorus; that is, in a counterfactual world in which 'Hesperus' and 'Phosphorus' were not used in the way that we use them, as names of this planet, but as names of some other objects, one could have had qualitatively identical evidence and concluded that 'Hesperus' and 'Phosphorus' named two different objects.[48] But we, using the names as we do right now, can say in advance, that if Hesperus and Phosphorus are one and the same, then in no other possible world can they be different. We use 'Hesperus' as the name of a certain body and 'Phosphorus' as the name of a certain body. We use them as names of those bodies in all possible worlds. If, in fact, they are the *same* body, then in any other possible world we have to use them as a name of that object. And so in any other possible world it will be true that Hesperus is Phosphorus. So two things are true: first, that we do not know *a priori* that Hesperus is Phosphorus, and are in no position to find out the answer except empirically. Second, this is so because we could have evidence qualitatively indistinguishable from the evidence we have and determine the reference of the two names by the positions of two planets in the sky, without the planets being the same.

[48] There is a more elaborate discussion of this point in the third lecture, where its relation to a certain sort of counterpart theory is also mentioned.

Of course, it is only a contingent truth (not true in every other possible world) that the star seen over there in the evening is the star seen over there in the morning, because there are possible worlds in which Phosphorus was not visible in the morning. But that contingent truth shouldn't be identified with the statement that Hesperus is Phosphorus. It could only be so identified if you thought that it was a necessary truth that Hesperus is visible over there in the evening or that Phosphorus is visible over there in the morning. But neither of those are necessary truths even if that's the way we pick out the planet. These are the contingent marks by which we identify a certain planet and give it a name.

PUTNAM, H.
IN: JOURNAL OF PHILOSOPHY 70 (1973)

MEANING AND REFERENCE *

UNCLEAR as it is, the traditional doctrine that the notion "meaning" possesses the extension/intension ambiguity has certain typical consequences. The doctrine that the meaning of a term is a concept carried the implication that meanings are mental entities. Frege, however, rebelled against this "psychologism." Feeling that meanings are *public* property—that the *same* meaning can be "grasped" by more than one person and

* To be presented in an APA symposium on Reference, December 28, 1973. Commentators will be Charles Chastain and Keith S. Donnellan; for Donnellan's paper, see this JOURNAL, this issue, 711–712; Professor Chastain's comments are not available at this time.

A very much expanded version of this paper will appear in volume 7 or 8 of Minnesota Studies in the Philosophy of Science (edited by Keith Gunderson), under the title "The Meaning of 'Meaning'."

by persons at different times—he identified concepts (and hence "intensions" or meanings) with abstract entities rather than mental entities. However, "grasping" these abstract entities was still an individual psychological act. None of these philosophers doubted that understanding a word (knowing its intension) was just a matter of being in a certain psychological state (somewhat in the way in which knowing how to factor numbers in one's head is just a matter of being in a certain very complex psychological state).

Secondly, the timeworn example of the two terms 'creature with a kidney' and 'creature with a heart' does show that two terms can have the same extension and yet differ in intension. But it was taken to be obvious that the reverse is impossible: two terms cannot differ in extension and have the same intension. Interestingly, no argument for this impossibility was ever offered. Probably it reflects the tradition of the ancient and medieval philosophers, who assumed that the concept corresponding to a term was just a conjunction of predicates, and hence that the concept corresponding to a term must *always* provide a necessary and sufficient condition for falling into the extension of the term. For philosophers like Carnap, who accepted the verifiability theory of meaning, the concept corresponding to a term provided (in the ideal case, where the term had "complete meaning") a *criterion* for belonging to the extension (not just in the sense of "necessary and sufficient condition," but in the strong sense of *way of recognizing* whether a given thing falls into the extension or not). So theory of meaning came to rest on two unchallenged assumptions:

(1) That knowing the meaning of a term is just a matter of being in a certain psychological state (in the sense of "psychological state," in which states of memory and belief are "psychological states"; no one thought that knowing the meaning of a word was a continuous state of consciousness, of course).

(2) That the meaning of a term determines its extension (in the sense that sameness of intension entails sameness of extension).

I shall argue that these two assumptions are not jointly satisfied by *any* notion, let alone any notion of meaning. The traditional concept of meaning is a concept which rests on a false theory.

ARE MEANINGS IN THE HEAD?

For the purpose of the following science-fiction examples, we shall suppose that somewhere there is a planet we shall call Twin Earth. Twin Earth is very much like Earth: in fact, people on Twin Earth even speak *English*. In fact, apart from the differences we shall specify in our science-fiction examples, the reader may suppose

that Twin Earth is *exactly* like Earth. He may even suppose that he has a *Doppelganger*—an identical copy—on Twin Earth, if he wishes, although my stories will not depend on this.

Although some of the people on Twin Earth (say, those who call themselves "Americans" and those who call themselves "Canadians" and those who call themselves "Englishmen," etc.) speak English, there are, not surprisingly, a few tiny differences between the dialects of English spoken on Twin Earth and standard English.

One of the peculiarities of Twin Earth is that the liquid called "water" is not H_2O but a different liquid whose chemical formula is very long and complicated. I shall abbreviate this chemical formula simply as XYZ. I shall suppose that XYZ is indistinguishable from water at normal temperatures and pressures. Also, I shall suppose that the oceans and lakes and seas of Twin Earth contain XYZ and not water, that it rains XYZ on Twin Earth and not water, etc.

If a space ship from Earth ever visits Twin Earth, then the supposition at first will be that 'water' has the same meaning on Earth and on Twin Earth. This supposition will be corrected when it is discovered that "water" on Twin Earth is XYZ, and the Earthian space ship will report somewhat as follows.

"On Twin Earth the word 'water' means XYZ."

Symmetrically, if a space ship from Twin Earth ever visits Earth, then the supposition at first will be that the word 'water' has the same meaning on Twin Earth and on Earth. This supposition will be corrected when it is discovered that "water" on Earth is H_2O, and the Twin Earthian space ship will report:

"On Earth the word 'water' means H_2O."

Note that there is no problem about the extension of the term 'water': the word simply has two different meanings (as we say); in the sense in which it is used on Twin Earth, the sense of water_{TE}, what *we* call "water" simply isn't water, while in the sense in which it is used on Earth, the sense of water_{E}, what the Twin Earthians call "water" simple isn't water. The extension of 'water' in the sense of water_{E} is the set of all wholes consisting of H_2O molecules, or something like that; the extension of water in the sense of water_{TE} is the set of all wholes consisting of XYZ molecules, or something like that.

Now let us roll the time back to about 1750. The typical Earthian speaker of English did not know that water consisted of hydrogen and oxygen, and the typical Twin-Earthian speaker of English did

not know that "water" consisted of XYZ. Let $Oscar_1$ be such a typical Earthian English speaker, and let $Oscar_2$ be his counterpart on Twin Earth. You may suppose that there is no belief that $Oscar_1$ had about water that $Oscar_2$ did not have about "water." If you like, you may even suppose that $Oscar_1$ and $Oscar_2$ were exact duplicates in appearance, feelings, thoughts, interior monologue, etc. Yet the extension of the term 'water' was just as much H_2O on Earth in 1750 as in 1950; and the extension of the term 'water' was just as much XYZ on Twin Earth in 1750 as in 1950. $Oscar_1$ and $Oscar_2$ understood the term 'water' differently in 1750 *although they were in the same psychological state,* and although, given the state of science at the time, it would have taken their scientific communities about fifty years to discover that they understood the term 'water' differently. Thus the extension of the term 'water' (and, in fact, its "meaning" in the intuitive preanalytical usage of that term) is *not* a function of the psychological state of the speaker by itself.[1]

But, it might be objected, why should we accept it that the term 'water' had the same extension in 1750 and in 1950 (on both Earths)? Suppose I point to a glass of water and say "this liquid is called water." My "ostensive definition" of water has the following empirical presupposition: that the body of liquid I am pointing to bears a certain sameness relation (say, *x is the same liquid as y,* or *x is the* $same_L$ *as y*) to most of the stuff I and other speakers in my linguistic community have on other occasions called "water." If this presupposition is false because, say, I am—unknown to me—pointing to a glass of gin and not a glass of water, then I do not intend my ostensive definition to be accepted. Thus the ostensive definition conveys what might be called a "defeasible" necessary and sufficient condition: the necessary and sufficient condition for being water is bearing the relation $same_L$ to the stuff in the glass; but this is the necessary and sufficient condition only if the empirical presupposition is satisfied. If it is not satisfied, then one of a series of, so to speak, "fallback" conditions becomes activated.

The key point is that the relation $same_L$ is a *theoretical* relation: whether something is or is not the same liquid as *this* may take an indeterminate amount of scientific investigation to determine. Thus, the fact that an English speaker in 1750 might have called XYZ "water," whereas he or his successors would not have called XYZ water in 1800 or 1850 does not mean that the "meaning" of 'water' changed for the average speaker in the interval. In 1750

[1] See fn 2, p. 710 below, and the corresponding text.

or in 1850 or in 1950 one might have pointed to, say, the liquid in Lake Michigan as an example of "water." What changed was that in 1750 we would have mistakenly thought that XYZ bore the relation $same_L$ to the liquid in Lake Michigan, whereas in 1800 or 1850 we would have known that it did not.

Let us now modify our science-fiction story. I shall suppose that molybdenum pots and pans *can't* be distinguished from aluminum pots and pans save by an expert. (This could be true for all I know, and, *a fortiori,* it could be true for all I know by virtue of "knowing the meaning" of the words *aluminum* and *molybdenum.*) We will now suppose that molybdenum is as common on Twin Earth as aluminum is on Earth, and that aluminum is as rare on Twin Earth as molybdenum is on Earth. In particular, we shall assume that "aluminum" pots and pans are made of molybdenum on Twin Earth. Finally, we shall assume that the words 'aluminum' and 'molybdenum' are *switched* on Twin Earth: 'aluminum' is the name of *molybdenum,* and 'molybdenum' is the name of *aluminum.* If a space ship from Earth visited Twin Earth, the visitors from Earth probably would not suspect that the "aluminum" pots and pans on Twin Earth were not made of aluminum, especially when the Twin Earthians *said* they were. But there is one important difference between the two cases. An Earthian metallurgist could tell very easily that "aluminum" was molybdenum, and a Twin Earthian metallurgist could tell equally easily that aluminum was "molybdenum." (The shudder quotes in the preceding sentence indicate Twin Earthian usages.) Whereas in 1750 no one on either Earth or Twin Earth could have distinguished water from "water," the confusion of aluminum with "aluminum" involves only a part of the linguistic communities involved.

This example makes the same point as the preceding example. If $Oscar_1$ and $Oscar_2$ are standard speakers of Earthian English and Twin Earthian English, respectively, and neither is chemically or metallurgically sophisticated, then there may be no difference at all in their psychological states when they use the word 'aluminum'; nevertheless, we have to say that 'aluminum' has the extension *aluminum* in the idiolect of $Oscar_1$ and the extension *molybdenum* in the idiolect of $Oscar_2$. (Also we have to say that $Oscar_1$ and $Oscar_2$ mean different things by 'aluminum'; that 'aluminum' has a different meaning on Earth than it does on Twin Earth, etc.) Again we see that the psychological state of the speaker does *not* determine the extension (*or* the "meaning," speaking preanalytically) of the word.

Before discussing this example further, let me introduce a *non*-science-fiction example. Suppose you are like me and cannot tell an elm from a beech tree. We still say that the extension of 'elm' in my idiolect is the same as the extension of 'elm' in anyone else's, viz., the set of all elm trees, and that the set of all beech trees is the extension of 'beech' in *both* of our idiolects. Thus 'elm' in my idiolect has a different extension from 'beech' in your idiolect (as it should). Is it really credible that this difference in extension is brought about by some difference in our *concepts*? My *concept* of an elm tree is exactly the same as my concept of a beech tree (I blush to confess). If someone heroically attempts to maintain that the difference between the extension of 'elm' and the extension of 'beech' in *my* idiolect is explained by a difference in my psychological state, then we can always refute him by constructing a "Twin Earth" example—just let the words 'elm' and 'beech' be switched on Twin Earth (the way 'aluminum' and "molybdenum' were in the previous example). Moreover, suppose I have a *Doppelganger* on Twin Earth who is molecule for molecule "identical" with me. If you are a dualist, then also suppose my Doppelganger thinks the same verbalized thoughts I do, has the same sense data, the same dispositions, etc. It is absurd to think *his* psychological state is one bit different from mine: yet he "means" *beech* when he says "elm," and *I* "mean" *elm* when I say "elm." Cut the pie any way you like, "meanings" just ain't in the *head!*

A SOCIOLINGUISTIC HYPOTHESIS

The last two examples depend upon a fact about language that seems, surprisingly, never to have been pointed out: that there is *division of linguistic labor*. We could hardly use such words as 'elm' and 'aluminum' if no one possessed a way of recognizing elm trees and aluminum metal; but not everyone to whom the distinction is important has to be able to make the distinction. Let us shift the example; consider *gold*. Gold is important for many reasons: it is a precious metal; it is a monetary metal; it has symbolic value (it is important to most people that the "gold" wedding ring they wear *really* consist of gold and not just *look* gold); etc. Consider our community as a "factory": in this "factory" some people have the "job" of *wearing gold wedding rings;* other people have the "job" of selling gold wedding rings; still other people have the job of *telling whether or not something is really gold*. It is not at all necessary or efficient that every one who wears a gold ring (or a gold cufflink, etc.), or discusses the "gold standard," etc., engage in buying and selling gold. Nor is it necessary or efficient that every one who buys and sells gold be able to tell whether or not something is really gold in a society where this form of dishonesty is uncommon (selling fake gold) and in which one can easily consult an expert in case of doubt. And it is *certainly* not necessary or efficient that every one who has occasion to buy or wear gold be able to tell with any reliability whether or not something is really gold.

The foregoing facts are just examples of mundane division of labor (in a wide sense). But they engender a division of linguistic labor: every one to whom gold is important for any reason has to *acquire* the word 'gold'; but he does not have to acquire the *method of recognizing* whether something is or is not gold. He can rely on a special subclass of speakers. The features that are generally thought to be present in connection with a general name—necessary and sufficient conditions for membership in the extension, ways of recognizing whether something is in the extension, etc.—are all present in the linguistic community *considered as a collective body;* but that collective body divides the "labor" of knowing and employing these various parts of the "meaning" of 'gold'.

This division of linguistic labor rests upon and presupposes the division of *non*linguistic labor, of course. If only the people who know how to tell whether some metal is really gold or not have any reason to have the word 'gold' in their vocabulary, then the word 'gold' will be as the word 'water' was in 1750 with respect to that subclass of speakers, and the other speakers just won't acquire it at all. And some words do not exhibit any division of linguistic labor: 'chair', for example. But with the increase of division of labor in the society and the rise of science, more and more words begin to exhibit this kind of division of labor. 'Water', for example, did not exhibit it at all before the rise of chemistry. Today it is obviously necessary for every speaker to be able to recognize water (reliably under normal conditions), and probably most adult speakers even know the necessary and sufficient condition "water is H_2O," but only a few adult speakers could distinguish water from liquids that superficially resembled water. In case of doubt, other speakers would rely on the judgment of these "expert" speakers. Thus the way of recognizing possessed by these "expert" speakers is also, through them, possessed by the collective linguistic body, even though it is not possessed by each individual member of the body, and in this way the most *recherché* fact about water may become part of the *social* meaning of the word although unknown to almost all speakers who acquire the word.

It seems to me that this phenomenon of division of linguistic

labor is one that it will be very important for sociolinguistics to investigate. In connection with it, I should like to propose the following hypothesis:

HYPOTHESIS OF THE UNIVERSALITY OF THE DIVISION OF LINGUISTIC LABOR: Every linguistic community exemplifies the sort of division of linguistic labor just described; that is, it possesses at least some terms whose associated "criteria" are known only to a subset of the speakers who acquire the terms, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets.

It is easy to see how this phenomenon accounts for some of the examples given above of the failure of the assumptions (1 and 2). When a term is subject to the division of linguistic labor, the "average" speaker who acquires it does not acquire anything that fixes its extension. In particular, his individual psychological state *certainly* does not fix its extension; it is only the sociolinguistic state of the collective linguistic body to which the speaker belongs that fixes the extension.

We may summarize this discussion by pointing out that there are two sorts of tools in the world: there are tools like a hammer or a screwdriver which can be used by one person; and there are tools like a steamship which require the cooperative activity of a number of persons to use. Words have been thought of too much on the model of the first sort of tool.

INDEXICALITY AND RIGIDITY

The first of our science-fiction examples—'water' on Earth and on Twin Earth in 1750—does not involve division of linguistic labor, or at least does not involve it in the same way the examples of 'aluminum' and 'elm' do. There were not (in our story, anyway) any "experts" on water on Earth in 1750, nor any experts on "water" on Twin Earth. The example *does* involve things which are of fundamental importance to the theory of reference and also to the theory of necessary truth, which we shall now discuss.

Let W_1 and W_2 be two possible worlds in which I exist and in which this glass exists and in which I am giving a meaning explanation by pointing to this glass and saying "This is water." Let us suppose that in W_1 the glass is full of H_2O and in W_2 the glass is full of XYZ. We shall also suppose that W_1 is the *actual* world, and that XYZ is the stuff typically called "water" in the world W_2 (so that the relation between English speakers in W_1 and English speakers in W_2 is exactly the same as the relation between English speakers on Earth and English speakers on Twin Earth).

Then there are two theories one might have concerning the meaning of 'water':

(1) One might hold that 'water' was *world-relative* but *constant* in meaning (i.e., the word has a constant relative meaning). On this theory, 'water' means the same in W_1 and W_2; it's just that water is H_2O in W_1, and water is XYZ in W_2.

(2) One might hold that water is H_2O in all worlds (the stuff called "water" in W_2 isn't water), but 'water' doesn't have the same meaning in W_1 and W_2.

If what was said before about the Twin Earth case was correct, then (2) is clearly the correct theory. When I say "*this* (liquid) is water," the "this" is, so to speak, a *de re* "this"—i.e., the force of my explanation is that "water" is whatever bears a certain equivalence relation (the relation we called "$same_L$" above) to the piece of liquid referred to as "this" *in the actual world.*

We might symbolize the difference between the two theories as a "scope" difference in the following way. On theory (1), the following is true:

(1′) (For every world W) (For every x in W) (x is water $\equiv$ x bears $same_L$ to the entity referred to as "this" in W)

while on theory (2):

(2′) (For every world W) (For every x in W) (x is water $\equiv$ x bears $same_L$ to the entity referred to as "this" *in the actual world* W_1)

I call this a "scope" difference because in (1′) 'the entity referred to as "this"' is within the scope of 'For every world W'—as the qualifying phrase 'in W' makes explicit—whereas in (2′) 'the entity referred to as "this"' means "the entity referred to as 'this' *in the actual world,*" and has thus a reference *independent* of the bound variable 'W'.

Kripke calls a designator "rigid" (in a given sentence) if (in that sentence) it refers to the same individual in every possible world in which the designator designates. If we extend this notion of rigidity to substance names, then we may express Kripke's theory and mine by saying that the term 'water' is *rigid.*

The rigidity of the term 'water' follows from the fact that when I give the "ostensive definition": "*this* (liquid) is water," I intend (2′) and not (1′).

We may also say, following Kripke, that when I give the ostensive definition "*this* (liquid) is water," the demonstrative 'this' is *rigid.*

What Kripke was the first to observe is that this theory of the meaning (or "use," or whatever) of the word 'water' (and other natural-kind terms as well) has startling consequences for the theory of necessary truth.

To explain this, let me introduce the notion of a *cross-world relation.* A two-term relation R will be called *cross-world* when it is understood in such a way that its extension is a set of ordered pairs of individuals *not all in the same possible world.* For example, it is easy to understand the relation *same height as* as a cross-world relation: just understand it so that, e.g., if x is an individual in a world W_1 who is 5 feet tall (in W_1) and y is an individual in W_2 who is 5 feet tall (in W_2), then the ordered pair x,y belongs to the extension of *same height as.* (Since an individual may have different heights in different possible worlds in which that same individual exists, strictly speaking, it is not the ordered pair x,y that constitutes an element of the extension of *same height as,* but rather the ordered pair *x-in-world-W_1, y-in-world-W_2.*)

Similarly, we can understand the relation $same_L$ (same liquid as) as a cross-world relation by understanding it so that a liquid in world W_1 which has the same important physical properties (in W_1) that a liquid in W_2 possesses (in W_2) bears $same_L$ to the latter liquid.

Then the theory we have been presenting may be summarized by saying that an entity x, in an arbitrary possible world, is *water* if and only if it bears the relation $same_L$ (construed as a cross-world relation) to the stuff *we* call "water" in the actual world.

Suppose, now, that I have not yet discovered what the important physical properties of water are (in the actual world)—i.e., I don't yet know that water is H_2O. I may have ways of *recognizing* water that are successful (of course, I may make a small number of mistakes that I won't be able to detect until a later stage in our scientific development), but not know the microstructure of water. If I agree that a liquid with the superficial properties of "water" but a different microstructure *isn't really water,* then my ways of recognizing water cannot be regarded as an analytical specification of what *it is to be* water. Rather, the operational definition, like the ostensive one, is simply a way of pointing out a standard—pointing out the stuff *in the actual world* such that, for x to be water, in *any* world, is for x to bear the relation $same_L$ to the *normal* members of the class of *local* entities that satisfy the operational definition. "Water" on Twin Earth is not water, even if it satisfies the operational definition, because it doesn't bear $same_L$ to the *local* stuff that satisfies the operational definition, and local stuff that satisfies the operational definition but has a microstructure different from the rest of the local stuff that satisfies the operational definition isn't water either, because it doesn't bear $same_L$ to the *normal* examples of the local "water."

Suppose, now, that I discover the microstructure of water—that water is H_2O. At this point I will be able to say that the stuff on Twin Earth that I earlier *mistook* for water isn't really water. In the same way, if you describe, not another planet in the actual universe, but another possible universe in which there is stuff with the chemical formula XYZ which passes the "operational test" for *water,* we shall have to say that that stuff isn't water but merely XYZ. You will not have described a possible world in which "water is XYZ," but merely a possible world in which there are lakes of XYZ, people drink XYZ (and not water), or whatever. In fact, once we have discovered the nature of water, nothing counts as a possible world in which water doesn't have that nature. Once we have discovered that water (in the actual world) is H_2O, *nothing counts as a possible world in which water isn't H_2O.*

On the other hand, we can perfectly well imagine having experiences that would convince us (and that would make it rational to believe that) water *isn't* H_2O. In that sense, it is conceivable that water isn't H_2O. It is conceivable but it isn't possible! Conceivability is no proof of possibility.

Kripke refers to statements that are rationally unrevisable (assuming there are such) as *epistemically necessary.* Statements that are true in all possible worlds he refers to simply as necessary (or sometimes as "metaphysically necessary"). In this terminology, the point just made can be restated as: a statement can be (metaphysically) necessary and epistemically contingent. Human intuition has no privileged access to metaphysical necessity.

In this paper, our interest is in theory of meaning, however, and not in theory of necessary truth. Words like 'now', 'this', 'here' have long been recognized to be *indexical,* or *token-reflexive*—i.e., to have an extension which varies from context to context or token to token. For these words, no one has ever suggested the traditional theory that "intension determines extension." To take our Twin Earth example: if I have a *Doppelganger* on Twin Earth, then when I think "I have a headache," *he* thinks "I have a headache." But the extension of the particular token of 'I' in his verbalized thought is himself (or his unit class, to be precise), while the extension of the token of 'I' in *my* verbalized thought is *me*

(or my unit class, to be precise). So the same word, 'I', has two different extensions in two different idiolects; but it does not follow that the concept I have of myself is in any way different from the concept my Doppelganger has of himself.

Now then, we have maintained that indexicality extends beyond the *obviously* indexical words and morphemes (e.g., the tenses of verbs). Our theory can be summarized as saying that words like 'water' have an unnoticed indexical component: "water" is stuff that bears a certain similarity relation to the water *around here*. Water at another time or in another place or even in another possible world has to bear the relation $same_L$ to *our* "water" *in order to be water*. Thus the theory that (1) words have "intensions," which are something like concepts associated with the words by speakers; and (2) intension determines extension—cannot be true of natural-kind words like 'water' for the same reason it cannot be true of obviously indexical words like 'I'.

The theory that natural-kind words like 'water' are indexical leaves it open, however, whether to say that 'water' in the Twin Earth dialect of English has the same *meaning* as 'water' in the Earth dialect and a different extension—which is what we normally say about 'I' in different idiolects—thereby giving up the doctrine that "meaning (intension) determines extension," or to say, as we have chosen to do, that difference in extension is *ipso facto* a difference in meaning for natural-kind words, thereby giving up the doctrine that meanings are concepts, or, indeed, mental entities of *any* kind.[2]

It should be clear, however, that Kripke's doctrine that natural-kind words are rigid designators and our doctrine that they are indexical are but two ways of making the same point.

We have now seen that the extension of a term is not fixed by a concept that the individual speaker has in his head, and this is true both because extension is, in general, determined *socially*—there is division of linguistic labor as much as of "real" labor—and

[2] Our reasons for rejecting the first option—to say that 'water' has the same meaning on Earth and on Twin Earth, while giving up the doctrine that meaning determines reference—are presented in "The Meaning of 'Meaning'." They may be illustrated thus: Suppose 'water' has the same meaning on Earth and on Twin Earth. Now, let the word 'water' become phonemically different on Twin Earth—say, it becomes 'quaxel'. Presumably, this is not a change in meaning per se, on any view. So 'water' and 'quaxel' have the same meaning (although they refer to different liquids). But this is highly counterintuitive. Why not say, then, that 'elm' in my idiolect has the same meaning as 'beech' in your idiolect, although they refer to different trees?

because extension is, in part, determined *indexically*. The extension of our terms depends upon the actual nature of the particular things that serve as paradigms, and this actual nature is not, in general, fully known to the speaker. Traditional semantic theory leaves out two contributions to the determination of reference—the contribution of society and the contribution of the real world; a better semantic theory must encompass both.

HILARY PUTNAM

Harvard University

QUINE, W.V.O.

IN THREE WAYS OF PARADOX AND OTHER ESSAYS

15

Three Grades of Modal Involvement

There are several closely interrelated operators, called *modal* operators, which are characteristic of modal logic. There are the operators of *necessity, possibility, impossibility, non-necessity.* Also there are the binary operators, or connectives, of *strict implication* and *strict equivalence.* These various operators are easily definable in terms of one another. Thus impossibility is necessity of the negation; possibility and non-necessity are the negations of impossibility and necessity; and strict implication and strict equivalence are necessity of the material conditional and biconditional. In a philosophical examination of modal logic we may therefore conveniently limit ourselves for the most part to a single modal operator, that of *necessity.* Whatever may be said about necessity may be said also, with easy and obvious adjustments, about the other modes.

There are three different degrees to which we may allow our logic, or semantics, to embrace the idea of necessity. The first or least degree of acceptance is this: necessity is expressed by a *semantical predicate* attributable to statements as notational forms—hence attachable to names of statements. We write, e.g.:

From the *Proceedings of the XIth International Congress of Philosophy,* Brussels, 1953, Volume 14 (Amsterdam: North-Holland Publishing Co.).

(1) Nec '9 > 5',

(2) Nec (Sturm's theorem),

(3) Nec 'Napoleon escaped from Elba',

in each case attaching the predicate 'Nec' to a noun, a singular term, which is a *name of* the statement which is affirmed to be necessary (or necessarily true). Of the above examples, (1) and (2) would presumably be regarded as true and (3) as false; for the necessity concerned in modal logic is generally conceived to be of a logical or a priori sort.

A second and more drastic degree in which the notion of necessity may be adopted is in the form of a *statement operator.* Here we have no longer a predicate, attaching to names of statements as in (1)–(3), but a logical operator 'nec', which attaches to statements themselves, in the manner of the negation sign. Under this usage, (1) and (3) would be rendered rather as:

(4) nec (9 > 5),

(5) nec (Napoleon escaped from Elba),

and (2) would be rendered by prefixing 'nec' to Sturm's actual theorem rather than to its name. Thus whereas 'Nec' is a predicate or verb, 'is necessary', which attaches to a noun to form a statement, 'nec' is rather an adverb, 'necessarily', which attaches to a statement to form a statement.

Finally the third and gravest degree is expression of necessity by a sentence operator. This is an extension of the second degree, and goes beyond it in allowing the attachment of 'nec' not only to statements but also to open sentences, such as '$x > 5$', preparatory to the ultimate attachment of quantifiers:

(6) (x) nec $(x > 5)$,

(7) $(\exists x)$ nec $(x > 5)$,

(8) $(x)[x = 9 . \supset$ nec $(x > 5)]$.

The example (6) would doubtless be rated as false, and perhaps (7) and (8) as true.

I shall be concerned in this paper to bring out the logical and philosophical significance of these three degrees of acceptance of a necessity device.

I

I call an occurrence of a singular term in a statement *purely referential*[1] (Frege: *gerade*[2]), if, roughly speaking, the term serves in that particular context simply to refer to its object. Occurrences within quotation are not in general referential; e.g., the statements:

(9) 'Cicero' contains six letters,

(10) '9 > 5' contains just three characters

say nothing about the statesman Cicero or the number 9. Frege's criterion for referential occurrence is substitutivity of identity. Since

(11) Tully = Cicero,

(12) the number of planets = 9,

whatever is true of Cicero is true *ipso facto* of Tully (these being one and the same) and whatever is true of 9 is true of the number of planets. If by putting 'Tully' for 'Cicero' or 'the number of planets' for '9' in a truth, e.g., (9) or (10), we come out with a falsehood:

(13) 'Tully' contains six letters,

(14) 'the number of planets > 5' contains just three characters,

we may be sure that the position on which the substitution was made was not purely referential.

(9) must not be confused with:

(15) Cicero has a six-letter name,

which *does* say something about the man Cicero, and—unlike (9)—remains true when the name 'Cicero' is supplanted by 'Tully'.

Taking a hint from Russell,[3] we may speak of a context as *referentially opaque* when, by putting a statement ϕ into that context, we can cause a purely referential occurrence in ϕ to be not purely referential in the whole context. E.g., the context:

'. . .' contains just three characters

[1] *From a Logical Point of View*, pp. 75f, 139ff, 145.
[2] "Über Sinn und Bedeutung."
[3] Whitehead and Russell, 2d ed., Vol. 1, Appendix C.

is referentially opaque; for, the occurrence of '9' in '9 > 5' is purely referential, but the occurence of '9' in (10) is not. Briefly, a context is referentially opaque if it can render a referential occurrence non-referential.

Quotation is the referentially opaque context par excellence. Intuitively, what occurs inside a referentially opaque context may be looked upon as an orthographic accident, without logical status, like the occurrence of 'cat' in 'cattle'. The quotational context ' '9 > 5' ' of the statement '9 > 5' has, perhaps, unlike the context 'cattle' of 'cat', a deceptively systematic air which tempts us to think of its parts as somehow logically germane. Insofar as this temptation exists, it is salutary to paraphrase quotations by the following expedient. We may adopt names for each of our letters and other characters, and Tarski's '⌢' to express concatenation. Then, instead of naming a notational form by putting that notational form itself bodily between quotation marks, we can name it by spelling it. E.g., since 'μ' is mu, 'ϵ' is epsilon, and 'ν' is nu, the word '$\mu\epsilon\nu$' is mu⌢epsilon⌢nu. Similarly the statement '9 > 5' is n⌢g⌢f, if we adopt the letters 'n', 'g', and 'f' as names of the characters '9', '>', and '5'. The example (10) can thus be transcribed as:

(16) n⌢g⌢f contains just three characters.

Here there is no non-referential occurrence of the numeral '9', for there is no occurrence of it all; and here there is no referentially opaque containment of one statement by another, because there is no contained statement at all. Paraphrasing (10) into (16), so as to get rid altogether of the opaquely contained statement '9 > 5', is like paraphrasing 'cattle' into 'kine' so as to rid it of the merely orthographic occurrence of the term 'cat'. Neither paraphrase is mandatory, but both are helpful when the irreferential occurrences draw undue attention.

An occurrence of a statement as a part of a longer statement is called *truth-functional* if, whenever we supplant the contained statement by another statement having the same truth value, the containing statement remains unchanged in truth value. Naturally one would not expect occurrences of statements within referentially opaque contexts, such as quotations, to be truth-functional. E.g., the truth (10) becomes false when the contained statement '9 > 5' is supplanted by another, 'Napoleon escaped

from Elba', which has the same truth value as '$9 > 5$'. Again the truth (1) is carried, by that same substitution, into the falsehood (3). One might not expect occurrences of statements within statements to be truth-functional, in general, even when the contexts are not referentially opaque; certainly not when the contexts are referentially opaque.

In mathematical logic, however, a policy of *extensionality* is widely espoused: a policy of admitting statements within statements truth-functionally only (apart of course from such contexts as quotation, which are referentially opaque). Note that the semantical predicate 'Nec' as of (1)–(3) is reconcilable with this policy of extensionality, since whatever breach of extensionality it *prima facie* involves is shared by examples like (10) and attributable to the referential opacity of quotation. We can always switch to the spelling expedient, thus rewriting (1) as:

(17) Nec (n⌢g⌢f).

(17), like (16) and indeed (2) and unlike (1) and (3), contains no component statement but only a name of a statement.

The statement operator 'nec', on the other hand, is a premeditated departure from extensionality. The occurrence of the truth '$9 > 5$' in (4) is non-truth-functional, since by supplanting it by a different truth we can turn the true context (4) into a falsehood such as (5). Such occurrences, moreover, are not looked upon as somehow spurious or irrelevant to logical structure, like occurrences in quotation or like 'cat' in 'cattle'. On the contrary, the modal logic typified in (4) is usually put forward as a corrective of extensionality, a needed supplementation of an otherwise impoverished logic. Truth-functional occurrence is by no means the rule in ordinary language, as witness occurrences of statements governed by 'because', 'thinks that', 'wishes that', etc., as well as 'necessarily'. Modal logicians, adopting 'nec', have seen no reason to suppose that an adequate logic might adhere to a policy of extensionality.

But, for all the willingness of modal logicians to flout the policy of extensionality, is there really any difference—on the score of extensionality—between their statement operator 'nec' and the extensionally quite admissible semantical predicate 'Nec'? The latter was excusable, within a policy of extensionality, by citing the referential opacity of quotation. But the statement operator 'nec' is likewise excusable, within a policy of extensionality, by citing the referential opacity of 'nec' itself! To see the referential opacity of 'nec' we have only to note that (4) and (12) are true and yet this is false:

(18) nec (the number of planets > 5).

The statement operator 'nec' is, in short, on a par with quotation. (1) happens to be written with quotation marks and (4) without, but from the point of view of a policy of extensionality one is no worse than the other. (1) might be preferable to (4) only on the score of a possible ancillary policy of trying to reduce referentially opaque contexts to uniformly quotational form.

Genuine violation of the extensionality policy, by admitting non-truth-functional occurrences of statements within statements *without* referential opacity, is less easy than one at first supposes. Extensionality does not merely recommend itself on the score of simplicity and convenience; it rests on somewhat more compelling grounds, as the following argument will reveal. Think of 'p' as short for some statement, and think of '$F(p)$' as short for some containing true statement, such that the context represented by 'F' is not referentially opaque. Suppose further that the context represented by 'F' is such that logical equivalents are interchangeable, within it, *salvâ veritate*. (This is true in particular of 'nec'.) What I shall show is that the occurrence of 'p' in '$F(p)$' is then truth-functional. I.e., think of 'q' as short for some statement having the same truth value as 'p'; I shall show that '$F(q)$' is, like '$F(p)$', true.

What 'p' represents is a statement, hence true or false (and devoid of free 'x'). If 'p' is true, then the conjunction '$x = \Lambda \,.\, p$' is true of one and only one object x, viz., the empty class Λ; whereas if 'p' is false the conjunction '$x = \Lambda \,.\, p$' is true of no object x whatever. The class $\hat{x}(x = \Lambda \,.\, p)$, therefore, is the unit class $\iota\Lambda$ or Λ itself according as 'p' is true or false. Moreover, the equation:

$$\hat{x}(x = \Lambda \,.\, p) = \iota\Lambda$$

is, by the above considerations, *logically* equivalent to 'p'. Then, since '$F(p)$' is true and logical equivalents are interchangeable within it, this will be true:

(19) $F[\hat{x}(x = \Lambda \,.\, p) = \iota\Lambda]$.

Since 'p' and 'q' are alike in truth value, the classes $\hat{x}(x = \Lambda \,.\, p)$ and $\hat{x}(x = \Lambda \,.\, q)$ are both $\iota\Lambda$ or both Λ; so

$$\hat{x}(x = \Lambda \,.\, p) = \hat{x}(x = \Lambda \,.\, q). \tag{20}$$

Since the context represented by 'F' is not referentially opaque, the occurrence of '$\hat{x}(x = \Lambda \,.\, p)$' in (19) is a purely referential occurrence and hence subject to the substitutivity of identity; so from (19) by (20) we can conclude that

$$F[\hat{x}(x = \Lambda \,.\, q) = \iota\Lambda].$$

Thence in turn, by the logical equivalence of '$\hat{x}(x = \Lambda \,.\, q) = \iota\Lambda$' to '$q$', we conclude that $F(q)$.

The above argument cannot be evaded by denying (20), as long as the notation in (20) is construed, as usual, as referring to classes. For classes, properly so-called, are one and the same if their members are the same—regardless of whether that sameness be a matter of logical proof or of historical accident. But the argument could be contested by one who does not admit class names '$\hat{x}(\ \ldots\ )$'. It could also be contested by one who, though admitting such class names, does not see a final criterion of referential occurrence in the substitutivity of identity, as applied to constant singular terms. These points will come up, perforce, when we turn to 'nec' as a sentence operator under quantification. Meanwhile the above argument does serve to show that the policy of extensionality has more behind it than its obvious simplicity and convenience, and that any real departure from the policy (at least where logical equivalents remain interchangeable) must involve revisions of the logic of singular terms.

The simpler earlier argument for the referential opacity of the statement operator 'nec', viz., observation of the truths (4) and (12) and the falsehood (18), could likewise be contested by one who either repudiates constant singular terms or questions the criterion of referential opacity which involves them. Short of adopting 'nec' as a full-fledged *sentence* operator, however, no such searching revisions of classical mathematical logic are required. We can keep to a classical theory of classes and singular terms, and even to a policy of extensionality. We have only to recognize, in the *statement* operator 'nec', a referentially opaque context comparable to the thoroughly legitimate and very convenient context of quotation. We can even look upon (4) and (5) as elliptical renderings of (1) and (3).

II

Something very much to the purpose of the semantical predicate 'Nec' is regularly needed in the theory of proof. When, e.g., we speak of the completeness of a deductive system of quantification theory, we have in mind some concept of *validity* as norm with which to compare the class of obtainable theorems. The notion of validity in such contexts is not identifiable with truth. A true statement is not a valid statement of quantification theory unless not only it but all other statements similar to it in quantificational structure are true. Definition of such a notion of validity presents no problem, and the importance of the notion for proof theory is incontestable.

A conspicuous derivative of the notion of quantificational validity is that of quantificational implication. One statement quantificationally implies another if the material conditional composed of the two statements is valid for quantification theory.

This reference to quantification theory is only illustrative. There are parallels for truth-function theory: a statement is valid for truth-function theory if it and all statements like it in truth-functional structure are true, and one statement truth-functionally implies another if the material conditional formed of the two statements is valid for truth-function theory.

And there are parallels, again, for logic taken as a whole: a statement is logically valid if it and all statements like it in logical structure are true, and one statement logically implies another if the material conditional formed of the two statements is logically valid.

Modal logic received special impetus years ago from a confused reading of '$\supset$', the material 'if–then', as 'implies': a confusion of the material conditional with the relation of implication.[4] Properly, whereas '$\supset$' or 'if–then' connects statements, 'implies' is a verb which connects names *of* statements and thus expresses a relation of the named statements. Carelessness over the distinction of use and mention having allowed this intrusion of 'implies' as a reading of '$\supset$', the protest thereupon arose that '$\supset$' in its material sense was too weak to do justice to 'implies', which connotes some-

[4] Notably in Whitehead and Russell.

thing like logical implication. Accordingly an effort was made to repair the discrepancy by introducing an improved substitute for '$\supset$', written '$\dashv\!\!\supset$' and called strict implication.[5] The initial failure to distinguish use from mention persisted; so $\dashv\!\!\supset$', though read 'implies' and motivated by the connotations of the word 'implies', functioned actually not as a verb but as a statement connective, a much strengthened 'if–then'. Finally, in recognition of the fact that logical implication is validity of the material conditional, a validity operator 'nec' was adopted to implement the definition of '$p \dashv\!\!\supset q$' as 'nec $(p \supset q)$'. Since '$\dashv\!\!\supset$' had been left at the level of a statement connective, 'nec' in turn was of course rendered as an operator directly attachable to statements—whereas 'is valid', properly, is a verb attachable to a name of a statement and expressing an attribute of the statement named.[6]

In any event, the use of 'nec' as statement operator is easily converted into use of 'Nec' as semantical predicate. We have merely to supply quotation marks, thus rewriting (4) and (5) as (1) and (3). The strong 'if–then', '$\dashv\!\!\supset$', can correspondingly be rectified to a relation of implication properly so-called. What had been:

(21) the witness lied $\dashv\!\!\supset$. the witness lied $\vee$ the owner is liable,

explained as:

(22) nec (the witness lied $\supset$. the witness lied $\vee$ the owner is liable),

becomes:

(23)

'the witness lied' implies 'the witness lied $\vee$ the owner is liable',

explained as:

(24)

Nec 'the witness lied $\supset$. the witness lied $\vee$ the owner is liable'.

Typically, in modal logic, laws are expressed with help of schematic letters 'p', 'q', etc., thus:

(25) $p \dashv\!\!\supset . p \vee q,$

(26) nec $(p \supset . p \vee q).$

[5] Lewis, *A Survey of Symbolic Logic*, Chap. 5.

[6] On the concerns of this paragraph and the next, see also §69 of Carnap, *Logical Syntax*, and §5 of my *Mathematical Logic*.

The schematic letters are to be thought of as supplanted by any specific statements so as to yield actual cases like (21) and (22). Now just as (21) and (22) are translatable into (23) and (24), so the schemata (25) and (26) themselves might be supposed translatable as:

(27) 'p' implies '$p \vee q$',

(28) Nec '$p \supset . p \vee q$'.

Here, however, we must beware of a subtle confusion. A quotation names precisely the expression inside it; a quoted 'p' names the sixteenth letter of the alphabet and nothing else. Thus whereas (25) and (26) are schemata or diagrams which depict the forms of actual statements, such as (21) and (22), on the other hand (27) and (28) are *not* schemata depicting the forms of actual statements such as (23) and (24). On the contrary, (27) and (28) are not schemata at all, but actual statements: statements *about* the specific schemata 'p', '$p \vee q$', and '$p \supset . p \vee q$' (with just those letters). Moreover, the predicates 'implies' and 'Nec' have thus far been looked upon as true only of statements, not of schemata; so in (27) and (28) they are misapplied (pending some deliberate extension of usage).

The letters 'p' and 'q' in (25) and (26) stand in place of statements. For translation of (25) and (26) into semantical form, on the other hand, we need some special variables which refer *to* statements and thus stand in place of names of statements. Let us use 'ϕ', 'ψ', etc., for that purpose. Then the analogues of (25) and (26) in semantical form can be rendered:

(29) ϕ implies the alternation of ϕ and ψ,

(30) Nec (the conditional of ϕ with the alternation of ϕ and ψ).

We can condense (29) and (30) by use of a conventional notation which I have elsewhere[7] called *quasi-quotation*, thus:

(31) ϕ implies $\ulcorner \phi \vee \psi \urcorner$,

(32) Nec $\ulcorner \phi \supset . \phi \vee \psi \urcorner$.

The relationship between the modal logic of statement operators and the semantical approach, which was pretty simple and obvious when we compared (21)–(22) with (23)–(24), is thus seen to take on some slight measure of subtlety at the stage of

[7] *Mathematical Logic*, §6.

(25)–(26); these correspond not to (27)–(28) but to (31)–(32). It is schemata like (25)–(26), moreover, and not actual statements like (21)–(22), that fill the pages of works on modal logic. However, be that as it may, it is in actual statements such as (21)–(24) that the point of modal logic lies, and it is the comparison of (21)–(22) with (23)–(24) that reflects the true relationship between the use of statement operators and that of semantical predicates. Schemata such as (25)–(26) are mere heuristic devices, useful in expounding the theory of (21)–(22) and their like; and the heuristic devices which bear similarly on (23)–(24) are (31)–(32).

Seeing how modal statement operators can be converted into semantical predicates, one may of course just note the conversion as a principle and leave it undone in practice. But there are five reasons why it is important to note it in principle. One is that the inclination to condemn '$\supset$' unduly, through a wrong association of 'if–then' with 'implies', is thereby removed. A second reason is that it is at the semantical or proof-theoretic level, where we talk *about* expressions and their truth values under various substitutions, that we make clear and useful sense of logical validity; and it is logical validity that comes nearest to being a clear explication of 'Nec', taken as a semantical predicate. A third reason is that in using 'Nec' as a semantical predicate we flaunt a familiar reminder of referential opacity, in the form of quotation marks. A fourth reason is that the adoption of 'nec' as a statement operator tempts one to go a step further and use it as a sentence operator subject to quantification. The momentousness of this further step—whereof more anon—tends to be overlooked save as one expressly conceives of the 'nec', in its use as statement operator, as shorthand for the semantical usage.

A fifth reason has to do with iteration. Since 'nec' attaches to a statement and produces a statement, 'nec' can then be applied again. On the other hand 'Nec' attaches to a name and yields a statement, to which, therefore, it cannot be applied again. An iterated 'nec', e.g.:

(33) $\quad$ nec nec$(x)(x$ is red $\supset x$ is red),

can of course be translated by our regular procedure into semantical form thus:

(34) $\quad$ Nec 'Nec '$(x)(x$ is red $\supset x$ is red)' ',

and we are thereby reminded that 'Nec' can indeed be iterated if we insert new quotation marks as needed. But the fact remains that (34) is, in contrast with (33), an unlikely move. For, suppose we have made fair sense of 'Nec' as logical validity, relative say to the logic of truth functions, quantification, and perhaps classes. The statement:

(35) $\quad$ $(x)(x$ is red $\supset x$ is red),

then, is typical of the statements to which we would attribute such validity; so

(36) $\quad$ Nec '$(x)(x$ is red $\supset x$ is red)'.

The validity of (35) resides in the fact that (35) is true and so are all other statements with the same quantificational and truth-functional structure as (35). Thus it is that (36) is *true*. But if (36) in turn is also *valid*, it is valid only in an extended sense with which we are not likely to have been previously concerned: a sense involving not only quantificational and truth-functional structure but also the semantical structure, somehow, of quotation and 'Nec' itself.

Ordinarily we work in a metalanguage, as in (36), treating of an object language, exemplified by (35). We would not rise to (34) except in the rare case where we want to treat the metalanguage by means of itself, and want furthermore to extend the notion of validity beyond the semantics of logic to the semantics of semantics. When on the other hand the statement operator 'nec' is used, iteration as in (33) is the most natural of steps; and it is significant that in modal logic there has been some question as to just what might most suitably be postulated regarding such iteration.[8]

The iterations need not of course be consecutive. In the use of modal statement operators we are led also into complex iterations such as:

(37) $\quad$ $p \supset q . \supset . \sim q \supset \sim p,$

short for:

(38) $\quad$ nec [nec $(p \supset q) \supset$ nec $(\sim q \supset \sim p)$].

[8] Cf. Lewis and Langford, pp. 497ff.

Or, to take an actual example:

(39) $(x)(x \text{ has mass}) \supset (\exists x)(x \text{ has mass}) .\supset. \sim(\exists x)(x \text{ has mass}) \supset \sim(x)(x \text{ has mass}),$

(40) $\text{nec}\ \{\text{nec}\ [(x)(x \text{ has mass}) \supset (\exists x)(x \text{ has mass})] \supset \text{nec}\ [\sim(\exists x)(x \text{ has mass}) \supset \sim(x)(x \text{ has mass})]\}$.

In terms of semantical predicates the correspondents of (39) and (40) are:

(41) ' '$(x)(x$ has mass)' implies '$(\exists x)(x$ has mass)' ' implies ' '$\sim(\exists x)(x$ has mass)' implies '$\sim(x)(x$ has mass)' ',

(42) Nec 'Nec '$(x)(x$ has mass) $\supset$ $(\exists x)(x$ has mass)' $\supset$ Nec '$\sim(\exists x)(x$ has mass) $\supset$ $\sim(x)(x$ has mass)' '.

But (41)–(42), like (34), have singularly little interest or motivation when we think of necessity semantically.

It is important to note that we must not translate the schemata (37)–(38) into semantical form in the manner:

' 'p' implies 'q' ' implies, etc.

To do so would be to compound, to an altogether horrifying degree, the error noted earlier of equating (25)–(26) to (27)–(28). The analogues of (37)–(38) in semantical application should be rendered rather:

(43) $\ulcorner\phi$ implies $\psi\urcorner$ implies $\ulcorner\ulcorner\sim\psi\urcorner$ implies $\ulcorner\sim\phi\urcorner\urcorner$,

(44) Nec $\ulcorner$Nec $\ulcorner\phi \supset \psi\urcorner \supset$ Nec $\ulcorner\sim\psi \supset \sim\phi\urcorner\urcorner$,

subject to some special conventions governing the nesting of quasi-quotations. Such conventions would turn on certain subtle considerations which will not be entered upon here. Suffice it to recall that the sort of thing formulated in (33)–(34) and (37)–(44) is precisely the sort of thing we are likely to see least point in formulating when we think of necessity strictly as a semantical predicate rather than a statement operator. It is impressive and significant that *most* of modal logic (short of quantified modal logic, to which we shall soon turn) is taken up with iterated cases like (33) and (37)–(40) which would simply not recommend themselves to our attention if necessity were held to the status of a semantical predicate and not depressed to the level of a statement operator.

Our reflections have favored the semantical side immensely,

but they must not be allowed to obscure the fact that even as a semantical predicate necessity can raise grave questions. There is no difficulty as long as necessity is construed as validity relative say to the logic of truth functions and quantification and perhaps classes. If we think of arithmetic as reduced to class theory, then such validity covers also the truths of arithmetic. But one tends to include further territory still; cases such as 'No bachelor is married', whose truth is supposed to depend on "meanings of terms" or on "synonymy" (e.g., the synonymy of 'bachelor' and 'man not married'). The synonymy relation on which such cases depend is supposedly a narrower relation than that of the mere coextensiveness of terms, and it is not known to be amenable to any satisfactory analysis. In short, necessity in semantical application tends to be identified with what philosophers call analyticity; and analyticity, I have argued elsewhere,[9] is a pseudo-concept which philosophy would be better off without.

As long as necessity in semantical application is construed simply as explicit truth-functional validity, on the other hand, or quantificational validity, or set-theoretic validity, or validity of any other well-determined kind, the logic of the semantical necessity predicate is a significant and very central strand of proof theory. But it is not modal logic, even unquantified modal logic, as the latter ordinarily presents itself; for it is a remarkably meager thing, bereft of all the complexities which are encouraged by the use of 'nec' as a statement operator. It is unquantified modal logic minus all principles which, explicitly or implicitly (via '$\supset$', etc.), involve iteration of necessity; and plus, if we are literal-minded, a pair of quotation marks after each 'Nec'.

III

Having adopted the operator '$\sim$' of negation as applicable to statements, one applies it without second thought to open sentences as well: sentences containing free variables ripe for quantification. Thus we can write not only '$\sim$(Socrates is mortal)' but also '$\sim(x$ is mortal)', from which, by quantification

[9] "Two dogmas of empiricism."

and further negation, we have '$\sim(x)$ $\sim$(x is mortal)' or briefly '$(\exists x)$ (x is mortal)'. With negation this is as it should be. As long as 'nec' is used as a statement operator, on a par with negation, the analogous course suggests itself again: we write not only 'nec $(9 > 5)$' but also 'nec $(x > 5)$', from which by quantification we can form (6)–(8) and the like.

This step brings us to 'nec' as sentence operator, Given 'nec' as statement operator, the step is natural. Yet it is a drastic one, for it suddenly obstructs the earlier expedient of translation into terms of 'Nec' as semantical predicate. We can reconstrue (4) and (5) at will as (1) and (3), but we cannot reconstrue:

(45) nec $(x > 5)$

correspondingly as:

(46) Nec '$x > 5$'.

'Nec' has been understood up to now as a predicate true only of statements, whereas (46) attributes it rather to an open sentence and is thus trivially false, at least pending some deliberate extension of usage. More important, whereas (45) is an open sentence with free 'x', (46) has no corresponding generality; (46) is simply a statement *about* a specific open sentence. For, it must be remembered that '$x > 5$' in quotation marks is a name of the specific quoted expression, with fixed letter 'x'. The 'x' in (46) cannot be reached by a quantifier. To write:

(47) (x)(Nec '$x > 5$'), $(\exists x)$(Nec '$x > 5$')

is like writing:

(48) (x)(Socrates is mortal), $(\exists x)$(Socrates is mortal);

the quantifier is followed by no germane occurrence of its variable. In a word, necessity as sentence operator does not go over into terms of necessity as semantical predicate.

Moreover, acceptance of necessity as a sentence operator implies an attitude quite opposite to our earlier one (in §§I–II above), which was that 'nec' as statement operator is referentially opaque. For, one would clearly have no business quantifying into a referentially opaque context; witness (47) above. We can reasonably infer '$(\exists x)$ nec $(x > 5)$' from 'nec $(9 > 5)$' only if we regard the latter as telling us something about the *object* 9, a number, viz. that it necessarily exceeds 5. If 'nec $(\ldots > 5)$'

can turn out true or false "of" the number 9 depending merely on how that number is referred to (as the falsity of (18) suggests), then evidently 'nec $(x > 5)$' expresses no genuine condition on objects of any kind. If the occurrence of '9' in 'nec $(9 > 5)$' is not purely referential, then putting 'x' for '9' in 'nec $(9 > 5)$' makes no more sense than putting 'x' for 'nine' within the context 'canine'.

But isn't it settled by the truth of (4) and (12) and the falsity of (18) that the occurrence of '9' in question is irreferential, and more generally that 'nec' is referentially opaque, and hence that 'nec' as a sentence operator under quantifiers is a mistake? No, not if one is prepared to accede to certain pretty drastic departures, as we shall see.

Thus far we have tentatively condemned necessity as general sentence operator on the ground that 'nec' is referentially opaque. Its referential opacity has been shown by a breakdown in the operation of putting one constant singular term for another which names the same object. But it may justly be protested that constant singular terms are a notational accident, not needed at the level of primitive notation.

For it is well known that primitively nothing in the way of singular terms is needed except the variables of quantification themselves. Derivatively all manner of singular terms may be introduced by contextual definition in conformity with Russell's theory of singular descriptions. Class names, in particular, which figured in the general argument for extensionality in §I above, may be got either by explaining '$\hat{x}(\ldots)$' as short for the contextually defined description '$(\iota y)(x)(x \in y . \equiv \ldots)$' or by adopting a separate set of contextual definitions for the purpose.[10]

Now the modal logician intent on quantifying into 'nec' sentences may say that 'nec' is not referentially opaque, but that it merely interferes somewhat with the contextual definition of singular terms. He may argue that '$(\exists x)$ nec $(x > 5)$' is not meaningless but true, and in particular that the number 9 is one of the things of which 'nec $(x > 5)$' is true. He may blame the real or apparent discrepancy in truth value between (4) and (18) simply on a queer behavior of contextually defined singular terms. Specifically he may hold that (18) is true if construed as:

[10]Cf. my *Methods of Logic*, §§36–38 (3d ed., §§41–43); *Mathematical Logic*, §§24, 26.

(49) $(\exists x)$[there are exactly x planets . nec $(x > 5)$]

and false if construed as:

(50) nec $(\exists x)$(there are exactly x planets . $x > 5$),

and that (18) as it stands is ambiguous for lack of a distinguishing mark favoring (49) or (50).[11] No such ambiguity arises in the contextual definition of a singular term in extensional logic (as long as the named object exists), and our modal logician may well deplore the complications which thus issue from the presence of 'nec' in his primitive notation. Still he can fairly protest that the erratic behavior of contextually defined singular terms is no reflection on the meaningfulness of his primitive notation, including his open 'nec' sentences and his quantification of them.

Looking upon quantification as fundamental, and constant singular terms as contextually defined, one must indeed concede the inconclusiveness of a criterion of referential opacity that rests on interchanges of constant singular terms. The objects of a theory are not properly describable as the things named by the singular terms; they are the values, rather, of the variables of quantification.[12] Fundamentally the proper criterion of referential opacity turns on quantification rather than naming, and is this: a referentially opaque context is one that cannot properly be *quantified into* (with quantifier outside the context and variable inside). Quotation, again, is the referentially opaque context par excellence; cf. (47). However, to object to necessity as sentence operator on the grounds of referential opacity so defined would be simply to beg the question.

Frege's criterion of referential occurrence, viz., substitutivity of identity, underlay the notion of referential opacity as developed in §I above. The statements of identity there concerned were formed of constant singular terms; cf. (11), (12). But there is a more fundamental form of the law of substitutivity of identity, which involves no constant singular terms, but only variables of quantification; viz.:

(51) $$(x)(y)(x = y \supset . Fx \equiv Fy).$$

This law is independent of any theory of singular terms, and cannot properly be challenged. For, to challenge it were simply to

[11] Thus Smullyan.

[12] See *From a Logical Point of View*, pp. 12ff, 75f, 102–110, 113ff, 148ff.

use the sign '=' in some unaccustomed way irrelevant to our inquiry. In any theory, whatever the shapes of its symbols, an open sentence whose free variables are 'x' and 'y' is an expression of identity only in case it fulfills (51) in the role of '$x = y$'. The generality of 'F' in (51) is this: 'Fx' is to be interpretable as any open sentence of the system in question, having 'x' as free (quantifiable) variable; and 'Fy', of course, is to be a corresponding context of 'y'.

If 'nec' is not referentially opaque, 'Fx' and 'Fy' in (51) can in particular be taken respectively as 'nec $(x = x)$' and 'nec $(x = y)$'. From (51), therefore, since surely 'nec $(x = x)$' is true for all x, we have:

(52) $$(x)(y)[x = y \supset \text{nec } (x = y)].$$

I.e., identity holds necessarily if it holds at all.

Let us not jump to the conclusion, just because (12) is true, that

(53) nec (the number of planets = 9).

This does not follow from (12) and (52) except with help of a law of universal instantiation, allowing us to put singular terms 'the number of planets' and '9' for the universally quantified 'x' and 'y' of (52). Such instantiation is allowable, certainly, in extensional logic; but it is a question of good behavior of constant singular terms, and we have lately observed that such behavior is not to be counted on when there is a 'nec' in the woodpile.

So our observations on necessity in quantificational application are, up to now, as follows. Necessity in such application is not *prima facie* absurd if we accept some interference in the contextual definition of singular terms. The effect of this interference is that constant singular terms cannot be manipulated with the customary freedom, even when their objects exist. In particular they cannot be used to instantiate universal quantifications, unless special supporting lemmas are at hand. A further effect of necessity in quantificational application is that objects come to be necessarily identical if identical at all.

There is yet a further consequence, and a particularly striking one: Aristotelian essentialism. This is the doctrine that some of the attributes of a thing (quite independently of the language in

which the thing is referred to, if at all) may be essential to the thing, and others accidental. E.g., a man, or talking animal, or featherless biped (for they are in fact all the same *things*), is essentially rational and accidentally two-legged and talkative, not merely qua man but qua itself. More formally, what Aristotelian essentialism says is that you can have open sentences—which I shall represent here as 'Fx' and 'Gx'—such that

(54) $$(\exists x)(\text{nec } Fx \,.\, Gx \,.\, \sim \text{nec } Gx).$$

An example of (54) related to the falsity of (53) might be:

$$(\exists x)[\text{nec}(x > 5) \,.\, \text{there are just } x \text{ planets} \,.\, \sim\text{nec (there are just } x \text{ planets)}],$$

such an object x being the number (by whatever name) which is variously known as 9 and the number of planets.

How Aristotelian essentialism as above formulated is required by quantified modal logic can be quickly shown. Actually something yet stronger can be shown: that there are open sentences 'Fx' and 'Gx' fulfilling not merely (54) but:

$$(x)(\text{nec } Fx \,.\, Gx \,.\, \sim \text{nec } Gx),$$

i.e.:

$$(x) \text{ nec } Fx \,.\, (x)\, Gx \,.\, (x) \sim \text{nec } Gx.$$

An appropriate choice of 'Fx' is easy: '$x = x$'. And an appropriate choice of 'Gx' is '$x = x \,.\, p$', where in place of 'p' any statement is chosen which is true but not necessarily true. Surely there *is* such a statement, for otherwise 'nec' would be a vacuous operator and there would be no point in modal logic.

Necessity as semantical predicate reflects a non-Aristotelian view of necessity: necessity resides in the way in which we say things, and not in the things we talk about. Necessity as statement operator is capable, we saw, of being reconstrued in terms of necessity as a semantical predicate, but has, nevertheless, its special dangers; it makes for an excessive and idle elaboration of laws of iterated modality, and it tempts one to a final plunge into quantified modality. This last complicates the logic of singular terms; worse, it leads us back into the metaphysical jungle of Aristotelian essentialism.

QUINE, W.V.O.
IN: JOURNAL OF PHILOSOPHY
53 (1956)

VOLUME LIII, No. 5 MARCH 1, 1956

THE JOURNAL OF PHILOSOPHY

QUANTIFIERS AND PROPOSITIONAL ATTITUDES[1]

I

THE incorrectness of rendering 'Ctesias is hunting unicorns' in the fashion:

$$(\exists x)(x \text{ is a unicorn} \,.\, \text{Ctesias is hunting } x)$$

is conveniently attested by the non-existence of unicorns, but is not due simply to that zoological lacuna. It would be equally incorrect to render 'Ernest is hunting lions' as:

(1) $$(\exists x)(x \text{ is a lion} \,.\, \text{Ernest is hunting } x),$$

where Ernest is a sportsman in Africa. The force of (1) is rather that there is some individual lion (or several) which Ernest is hunting; stray circus property, for example.

The contrast recurs in 'I want a sloop.' The version:

(2) $$(\exists x)(x \text{ is a sloop} \,.\, \text{I want } x)$$

is suitable insofar only as there may be said to be a certain sloop that I want. If what I seek is mere relief from slooplessness, then (2) conveys the wrong idea.

The contrast is that between what may be called the *relational* sense of lion-hunting or sloop-wanting, viz., (1)–(2), and the likelier or *notional* sense. Appreciation of the difference is evinced in Latin and Romance languages by a distinction of mood in subordinate clauses; thus 'Procuro un perro que habla' has the relational sense:

$$(\exists x)(x \text{ is a dog} \,.\, x \text{ talks} \,.\, \text{I seek } x)$$

as against the notional 'Procuro un perro que hable':

$$\text{I strive that } (\exists x)(x \text{ is a dog} \,.\, x \text{ talks} \,.\, \text{I find } x).$$

Pending considerations to the contrary in later pages, we may represent the contrast strikingly in terms of permutations of com-

[1] This paper sums up some points which I have set forth in various lectures at Harvard and Oxford from 1952 onward.

ponents. Thus (1) and (2) may be expanded (with some premeditated violence to both logic and grammar) thus:

(3) $(\exists x)$ (x is a lion . Ernest strives that Ernest finds x),

(4) $(\exists x)$ (x is a sloop . I wish that I have x),

whereas 'Ernest is hunting lions' and 'I want a sloop' in their notional senses may be rendered rather thus:

(5) Ernest strives that $(\exists x)$ (x is a lion . Ernest finds x),

(6) I wish that $(\exists x)$ (x is a sloop . I have x).

The contrasting versions (3)–(6) have been wrought by so paraphrasing 'hunt' and 'want' as to uncover the locutions 'strive that' and 'wish that,' expressive of what Russell has called *propositional attitudes*. Now of all examples of propositional attitudes, the first and foremost is *belief*; and, true to form, this example can be used to point up the contrast between relational and notional senses still better than (3)–(6) do. Consider the relational and notional senses of believing in spies:

(7) $(\exists x)$ (Ralph believes that x is a spy),

(8) Ralph believes that $(\exists x)$ (x is a spy).

Both may perhaps be ambiguously phrased as 'Ralph believes that someone is a spy,' but they may be unambiguously phrased respectively as 'There is someone whom Ralph believes to be a spy' and 'Ralph believes there are spies.' The difference is vast; indeed, if Ralph is like most of us, (8) is true and (7) false.

In moving over to propositional attitudes, as we did in (3)–(6), we gain not only the graphic structural contrast between (3)–(4) and (5)–(6) but also a certain generality. For, we can now multiply examples of striving and wishing, unrelated to hunting and wanting. Thus we get the relational and notional senses of wishing for a president:

(9) $(\exists x)$ (Witold wishes that x is president),

(10) Witold wishes that $(\exists x)$ (x is president).

According to (9), Witold has his candidate; according to (10) he merely wishes the appropriate form of government were in force. Also we open other propositional attitudes to similar consideration—as witness (7)–(8).

However, the suggested formulations of the relational senses—viz., (3), (4), (7), and (9)—all involve quantifying into a propositional-attitude idiom from outside. This is a dubious business, as may be seen from the following example.

There is a certain man in a brown hat whom Ralph has glimpsed several times under questionable circumstances on which we need not enter here; suffice it to say that Ralph suspects he is a spy. Also there is a gray-haired man, vaguely known to Ralph as rather a pillar of the community, whom Ralph is not aware of having seen except once at the beach. Now Ralph does not know it, but the men are one and the same. Can we say of this *man* (Bernard J. Ortcutt, to give him a name) that Ralph believes him to be a spy? If so, we find ourselves accepting a conjunction of the type:

(11) w sincerely denies '.....' . w believes that .....

as true, with one and the same sentence in both blanks. For, Ralph is ready enough to say, in all sincerity, 'Bernard J. Ortcutt is no spy.' If, on the other hand, with a view to disallowing situations of the type (11), we rule simultaneously that

(12) Ralph believes that the man in the brown hat is a spy,

(13) Ralph does not believe that the man seen at the beach is a spy,

then we cease to affirm any relationship between Ralph and any man at all. Both of the component 'that'-clauses are indeed about the man Ortcutt; but the 'that' must be viewed in (12) and (13) as sealing those clauses off, thereby rendering (12) and (13) compatible because not, as wholes, about Ortcutt at all. It then becomes improper to quantify as in (7); 'believes that' becomes, in a word, referentially opaque.[2]

No question arises over (8); it exhibits only a quantification *within* the 'believes that' context, not a quantification *into* it. What goes by the board, when we rule (12) and (13) both true, is just (7). Yet we are scarcely prepared to sacrifice the relational construction 'There is someone whom Ralph believes to be a spy,' which (7) as against (8) was supposed to reproduce.

The obvious next move is to try to make the best of our dilemma by distinguishing two senses of belief: $belief_1$, which disallows (11), and $belief_2$, which tolerates (11) but makes sense of (7). For $belief_1$, accordingly, we sustain (12)–(13) and ban (7) as nonsense. For $belief_2$, on the other hand, we sustain (7); and for *this* sense of belief we must reject (13) and acquiesce in the con-

[2] See *From a Logical Point of View* (Harvard University Press, 1953), pp. 142–159; also "Three Grades of Modal Involvement," *Proceedings of the Eleventh International Congress of Philosophy*, Vol. 14, pp. 65–81.

clusion that Ralph believes$_2$ that the man at the beach is a spy even though he *also* believes$_2$ (and believes$_1$) that the man at the beach is not a spy.

II

But there is a more suggestive treatment. Beginning with a single sense of belief, viz., belief$_1$ above, let us think of this at first as a relation between the believer and a certain *intension*, named by the 'that'-clause. Intensions are creatures of darkness, and I shall rejoice with the reader when they are exorcised, but first I want to make certain points with help of them. Now intensions named thus by 'that'-clauses, without free variables, I shall speak of more specifically as intensions of degree 0, or propositions. In addition I shall (for the moment) recognize intensions of degree 1, or attributes. These are to be named by prefixing a variable to a sentence in which it occurs free; thus z(*z* is a spy) is spyhood. Similarly we may specify intensions of higher degrees by prefixing multiple variables.

Now just as we have recognized a dyadic relation of belief between a believer and a proposition, thus:

(14) Ralph believes that Ortcutt is a spy,

so we may recognize also a triadic relation of belief among a believer, an object, and an attribute, thus:

(15) Ralph believes z(*z* is a spy) of Ortcutt.

For reasons which will appear, this is to be viewed not as dyadic belief between Ralph and the proposition *that* Ortcutt has z(*z* is a spy), but rather as an irreducibly triadic relation among the three things Ralph, z(*z* is a spy), and Ortcutt. Similarly there is tetradic belief:

(16) Tom believes yz(*y* denounced *z*) of Cicero and Catiline,

and so on.

Now we can clap on a hard and fast rule against quantifying into propositional-attitude idioms; but we give it the form now of a rule against quantifying into names of intensions. Thus, though (7) as it stands becomes unallowable, we can meet the needs which prompted (7) by quantifying rather into the triadic belief construction, thus:

(17) $(\exists x)$[Ralph believes z(*z* is a spy) of x].

Here then, in place of (7), is our new way of saying that there is someone whom Ralph believes to be a spy.

Belief$_1$ was belief so construed that a proposition might be believed when an object was specified in it in one way, and yet not believed when the same object was specified in another way; witness (12)–(13). Hereafter we can adhere uniformly to this narrow sense of belief, both for the dyadic case and for triadic and higher; in each case the term which names the intension (whether proposition or attribute or intension of higher degree) is to be looked on as referentially opaque.

The situation (11) is thus excluded. At the same time the effect of belief$_2$ can be gained, simply by ascending from dyadic to triadic belief as in (15). For (15) does relate the men Ralph and Ortcutt precisely as belief$_2$ was intended to do. (15) does remain true of Ortcutt under any designation; and hence the legitimacy of (17).

Similarly, whereas from:

Tom believes that Cicero denounced Catiline

we cannot conclude:

Tom believes that Tully denounced Catiline,

on the other hand we can conclude from:

Tom believes y(*y* denounced Catiline) of Cicero

that

Tom believes y(*y* denounced Catiline) of Tully,

and also that

(18) $(\exists x)$[Tom believes y(*y* denounced Catiline) of x].

From (16), similarly, we may infer that

(19) $(\exists w)(\exists x)$[Tom believes yz(*y* denounced *z*) of w and x].

Such quantifications as:

$(\exists x)$(Tom believes that x denounced Catiline),

$(\exists x)$[Tom believes y(*y* denounced x) of Cicero]

still count as nonsense, along with (7); but such legitimate purposes as these might have served are served by (17)–(19) and the like. Our names of intensions, and these only, are what count as referentially opaque.

Let us sum up our findings concerning the seven numbered statements about Ralph. (7) is now counted as nonsense, (8) as true, (12)–(13) as true, (14) as false, and (15) and (17) as true. Another that is true is:

(20) Ralph believes that the man seen at the beach is not a spy,

which of course must not be confused with (13).

The kind of exportation which leads from (14) to (15) should doubtless be viewed in general as implicative. Under the terms of our illustrative story, (14) happens to be false; but (20) is true, and it leads by exportation to:

(21) Ralph believes z(z is not a spy) of the man seen at the beach.

The man at the beach, hence Ortcutt, does not receive reference in (20), because of referential opacity; but he does in (21), so we may conclude from (21) that

(22) Ralph believes z(z is not a spy) of Ortcutt.

Thus (15) and (22) both count as true. This is not, however, to charge Ralph with contradictory beliefs. Such a charge might reasonably be read into:

(23) Ralph believes z(z is a spy . z is not a spy) of Ortcutt,

but this merely goes to show that it is undesirable to look upon (15) and (22) as implying (23).

It hardly needs be said that the barbarous usage illustrated in (15)–(19) and (21)–(23) is not urged as a practical reform. It is put forward by way of straightening out a theoretical difficulty, which, summed up, was as follows: Belief contexts are referentially opaque; therefore it is *prima facie* meaningless to quantify into them (at least with respect to persons or other extensional objects[3]); how then to provide for those indispensable relational statements of belief, like 'There is someone whom Ralph believes to be a spy'?

Let it not be supposed that the theory which we have been examining is just a matter of allowing unbridled quantification into belief contexts after all, with a legalistic change of notation. On the contrary, the crucial choice recurs at each point: quantify if you will, but pay the price of accepting situations of the type (11) with respect to each point at which you choose to quantify. In other words: distinguish as you please between referential and non-referential positions, but keep track, so as to treat each kind appropriately. The notation of intensions, of degree one and higher, is in effect a device for inking in a boundary between referential and non-referential occurrences of terms.

III

Striving and wishing, like believing, are propositional attitudes and referentially opaque. (3) and (4) are objectionable in the

[3] See *From a Logical Point of View*, pp. 150–154.

same way as (7), and our recent treatment of belief can be repeated for these propositional attitudes. Thus, just as (7) gave way to (17), so (3) and (4) give way to:

(24) $(\exists x)$[x is a lion . Ernest strives z(Ernest finds z) of x],

(25) $(\exists x)$[x is a sloop . I wish z(I have z) of x],

a certain breach of idiom being allowed for the sake of analogy in the case of 'strives.'

These examples came from a study of hunting and wanting. Observing in (3)–(4) the quantification into opaque contexts, then, we might have retreated to (1)–(2) and foreborne to paraphrase them into terms of striving and wishing. For (1)–(2) were quite straightforward renderings of lion-hunting and sloop-wanting in their relational senses; it was only the notional senses that really needed the breakdown into terms of striving and wishing, (5)–(6).

Actually, though, it would be myopic to leave the relational senses of lion-hunting and sloop-wanting at the unanalyzed stage (1)–(2). For, whether or not we choose to put these over into terms of wishing and striving, there are other relational cases of wishing and striving which require our consideration anyway —as witness (9). The untenable formulations (3)–(4) may indeed be either corrected as (24)–(25) or condensed back into (1)–(2); on the other hand we have no choice but to correct the untenable (9) on the pattern of (24)–(25), viz., as:

$(\exists x)$[Witold wishes y(y is president) of x].

The untenable versions (3)–(4) and (9) all had to do with wishing and striving in the relational sense. We see in contrast that (5)–(6) and (10), on the notional side of wishing and striving, are innocent of any illicit quantification into opaque contexts from outside. But now notice that exactly the same trouble begins also on the notional side, as soon as we try to say not just that Ernest hunts lions and I want a sloop, but that *someone* hunts lions or wants a sloop. This move carries us, ostensibly, from (5)–(6) to:

(26) $(\exists w)$[w strives that $(\exists x)$(x is a lion . w finds x)],

(27) $(\exists w)$[w wishes that $(\exists x)$(x is a sloop . w has x)],

and these do quantify unallowably into opaque contexts.

We know how, with help of the attribute apparatus, to put (26)–(27) in order; the pattern, indeed, is substantially before us in (24)–(25). Admissible versions are:

$(\exists w)[w$ strives $y(\exists x)(x$ is a lion . y finds $x)$ of $w]$,

$(\exists w)[w$ wishes $y(\exists x)(x$ is a sloop . y has $x)$ of $w]$,

or briefly:

(28) $(\exists w)[w$ strives $y(y$ finds a lion) of $w]$,

(29) $(\exists w)[w$ wishes $y(y$ has a sloop) of $w]$.

Such quantification of the subject of the propositional attitude can of course occur in belief as well; and, if the subject is mentioned in the belief itself, the above pattern is the one to use. Thus 'Someone believes he is Napoleon' must be rendered:

$(\exists w)[w$ believes $y(y =$ Napoleon) of $w]$.

For concreteness I have been discussing belief primarily, and two other propositional attitudes secondarily: striving and wishing. The treatment is, we see, closely parallel for the three; and it will pretty evidently carry over to other propositional attitudes as well—e.g., hope, fear, surprise. In all cases my concern is, of course, with a special technical aspect of the propositional attitudes: the problem of quantifying in.

IV

There are good reasons for being discontent with an analysis that leaves us with propositions, attributes, and the rest of the intensions. Intensions are less economical than extensions (truth values, classes, relations), in that they are more narrowly individuated. The principle of their individuation, moreover, is obscure.

Commonly logical equivalence is adopted as the principle of individuation of intensions. More explicitly: if S and S′ are any two sentences with $n(\geqq 0)$ free variables, the same in each, then the respective intensions which we name by putting the n variables (or 'that,' if $n = 0$) before S and S′ shall be one and the same intension if and only if S and S′ are logically equivalent. But the relevant concept of logical equivalence raises serious questions in turn.[4]

Worse, granted certain usual logical machinery (such as is available in *Principia Mathematica*), this principle of individuation can be shown to contradict itself. For I have proved elsewhere,[5] using machinery solely of *Principia*, that if logical equivalence is taken as a sufficient condition of identity of attributes then mere coextensiveness becomes a sufficient condition as well. But then it follows that logical equivalence is not a necessary condition; so the described principle of individuation contradicts itself.

[4] See "Two Dogmas of Empiricism," in *From a Logical Point of View*; also "Carnap and Logical Truth," in Paul Arthur Schilpp (editor), *The Philosophy of Rudolf Carnap*, Library of Living Philosophers, at press.

[5] At the end of "On Frege's Way Out," *Mind*, Vol. 64 (1955).

The champion of intensions can be trusted, in the face of this result, to abandon either that principle of individuation of intensions or some one of the principles from *Principia* which was used in the proof. The fact remains that the intensions are at best a pretty obscure lot.

Yet it is evident enough that we cannot, in the foregoing treatment of propositional attitudes, drop the intensions in favor of the corresponding extensions. Thus, to take a trivial example, consider 'w is hunting unicorns.' On the analogy of (29), it becomes:

w strives $y(y$ finds a unicorn) of w.

Correspondingly for the hunting of griffins. Hence, if anyone w is to hunt unicorns without hunting griffins, the attributes

$y(y$ finds a unicorn),
$y(y$ finds a griffin)

must be distinct. But the corresponding classes are identical, being empty. So it is indeed the attributes, and not the classes, that were needed in our formulation. The same moral could be drawn, though less briefly, without appeal to empty cases.

But there is a way of dodging the intensions which merits serious consideration. Instead of speaking of intensions we can speak of sentences, naming these by quotation. Instead of:

w believes that .....

we may say:

w believes-true '.....'.

Instead of:

(30) w believes $y(\ldots y\ldots)$ of x

we may say:

(31) w believes '$\ldots y\ldots$' satisfied by x.

The words 'believes satisfied by' here, like 'believes of' before, would be viewed as an irreducibly triadic predicate. A similar shift can be made in the case of the other propositional attitudes, of course, and in the tetradic and higher cases.

This semantical reformulation is not, of course, intended to suggest that the subject of the propositional attitude speaks the language of the quotation, or any language. We may treat a mouses's fear of a cat as his fearing true a certain English sentence. This is unnatural without being therefore wrong. It is a little like describing a prehistoric ocean current as clockwise.

How, where, and on what grounds to draw a boundary between those who believe or wish or strive that p, and those who do not quite believe or wish or strive that p, is undeniably a vague and obscure affair. However, if anyone does approve of speaking of belief of a proposition at all and of speaking of a proposition in turn as meant by a sentence, then certainly he cannot object to our semantical reformulation 'w believes-true S' on any special grounds of obscurity; for, 'w believes-true S' is explicitly definable in *his* terms as 'w believes the proposition meant by S.' Similarly for the semantical reformulation (31) of (30); similarly for the tetradic and higher cases; and similarly for wishing, striving, and other propositional attitudes.

Our semantical versions do involve a relativity to language, however, which must be made explicit. When we say that w believes-true S, we need to be able to say what language the sentence S is thought of as belonging to; not because w needs to understand S, but because S might by coincidence exist (as a linguistic form) with very different meanings in two languages.[6] Strictly, therefore, we should think of the dyadic "believes-true S' as expanded to a triadic 'w believes-true S in L'; and correspondingly for (31) and its suite.

As noted two paragraphs back, the semantical form of expression:

(32) w believes-true '.....' in L

can be explained in intensional terms, for persons who favor them, as:

(33) w believes the proposition meant by '.....' in L,

thus leaving no cause for protest on the score of relative clarity. Protest may still be heard, however, on a different score: (32) and (33), though equivalent to each other, are not strictly equivalent to the 'w believes that .....' which is our real concern. For, it is argued, in order to infer (33) we need not only the information about w which 'w believes that .....' provides, but also some extraneous information about the language L. Church[7] brings

[6] This point is made by Alonzo Church, "On Carnap's Analysis of Statements of Assertion and Belief," *Analysis*, Vol. 10 (1950), pp. 97–99.

[7] Op. cit., with an acknowledgment to Langford.

the point out by appeal to translations, substantially as follows. The respective statements:

w believes that there are unicorns,

w believes the proposition meant by 'There are unicorns' in English

go into German as:

(34) w glaubt, dass es Einhörne gibt,

(35) w glaubt diejenige Aussage, die „There are unicorns" auf Englisch bedeutet,

and clearly (34) does not provide enough information to enable a German ignorant of English to infer (35).

The same reasoning can be used to show that 'There are unicorns' is not strictly or analytically equivalent to:

'There are unicorns' is true in English.

Nor, indeed, was Tarski's truth paradigm intended to assert analytic equivalence. Similarly, then, for (32) in relation to 'w believes that .....'; a systematic agreement in truth value can be claimed, and no more. This limitation will prove of little moment to persons who share my skepticism about analyticity.

What I find more disturbing about the semantical versions, such as (32), is the need of dragging in the language concept at all. What is a language? What degree of fixity is supposed? When do we have one language and not two? The propositional attitudes are dim affairs to begin with, and it is a pity to have to add obscurity to obscurity by bringing in language variables too. Only let it not be supposed that any clarity is gained by restituting the intensions.

W. V. QUINE

HARVARD UNIVERSITY

IN:
MARGALIT
MEANING AND USE

SAUL A. KRIPKE

A PUZZLE ABOUT BELIEF

In this paper I will present a puzzle about names and belief. A moral or two will be drawn about some other arguments that have occasionally been advanced in this area, but my main thesis is a simple one: that the puzzle *is* a puzzle. And, as a corollary, that any account of belief must ultimately come to grips with it. Any speculation as to solutions can be deferred.

The first section of the paper gives the theoretical background in previous discussion, and in my own earlier work, that led me to consider the puzzle. The background is by no means necessary to *state* the puzzle: As a philosophical puzzle, it stands on its own, and I think its fundamental interest for the problem of belief goes beyond the background that engendered it. As I indicate in the third section, the problem really goes beyond beliefs expressed using names, to a far wider class of beliefs. Nevertheless, I think that the background illuminates the genesis of the puzzle, and it will enable me to draw one moral in the concluding section.

The second section states some general principles which underlie our general practice of reporting beliefs. These principles are stated in much more detail than is needed to comprehend the puzzle; and there are variant formulations of the principles that would do as well. Neither this section nor the first is necessary for an intuitive grasp of the central problem, discussed in the third section, though they may help with fine points of the discussion. The reader who wishes rapid access to the central problem could skim the first two sections lightly on a first reading.

In one sense the problem may strike some as no puzzle at all. For, in the situation to be envisaged, all the relevant facts can be described in *one* terminology without difficulty. But, in *another* terminology, the situation seems to be impossible to describe in a consistent way. This will become clearer later.

I. PRELIMINARIES: SUBSTITUTIVITY

In other writings,[1] I developed a view of proper names closer in many ways to the old Millian paradigm of naming than to the Fregean tradition which probably was dominant until recently. According to Mill, a proper name is, so to speak, *simply* a name. It *simply* refers to its bearer, and has no other

linguistic function. In particular, unlike a definite description, a name does not describe its bearer as possessing any special identifying properties.

The opposing Fregean view holds that to each proper name, a speaker of the language associates some property (or conjunction of properties) which determines its referent as the unique thing fulfilling the associated property (or properties). This property(ies) constitutes the 'sense' of the name. Presumably, if '. . .' is a proper name, the associated properties are those that the speaker would supply, if asked, "Who is '. . .'?" If he would answer ". . . is the man who ————," the properties filling the second blank are those that determine the reference of the name for the given speaker and constitute its 'sense.' Of course, given the name of a famous historical figure, individuals may give different, and equally correct, answers to the "Who is . . . ?" question. Some may identify Aristotle as the philosopher who taught Alexander the Great, others as the Stagirite philosopher who studied with Plato. For these two speakers, the sense of "Aristotle" will differ: in particular, speakers of the second kind, but not of the first kind, will regard "Aristotle, if he existed, was born in Stagira" as analytic.[2] Frege (and Russell)[3] concluded that, strictly speaking, different speakers of English (or German!) ordinarily use a name such as 'Aristotle' in different senses (though with the same reference). Differences in properties associated with such names, strictly speaking, yield different idiolects.[4]

Some later theorists in the Frege-Russellian tradition have found this consequence unattractive. So they have tried to modify the view by 'clustering' the sense of the name (e.g., Aristotle is the thing having the following long list of properties, or at any rate most of them), or, better for the present purpose, socializing it (what determines the reference of 'Aristotle' is some roughly specified set of *community-wide* beliefs about Aristotle).

One way to point up the contrast between the strict Millian view and Fregean views involves — if we permit ourselves this jargon — the notion of propositional content. If a strict Millian view is correct, and the linguistic function of a proper name is completely exhausted by the fact that it names its bearer, it would appear that proper names of the same thing are everywhere interchangeable not only *salva veritate* but even *salva significatione*: the proposition expressed by a sentence should remain the same no matter what name of the object it uses. Of course this will not be true if the names are 'mentioned' rather than 'used': " 'Cicero' has six letters" differs from " 'Tully' has six letters" in truth value, let alone in content. (The example, of course, is Quine's.) Let us confine ourselves at this stage to

simple sentences involving no connectives or other sources of intensionality. If Mill is completely right, not only should "Cicero was lazy" have the same *truth value* as "Tully was lazy," but the two sentences should express the same *proposition*, have the same content. Similarly "Cicero admired Tully," "Tully admired Cicero," "Cicero admired Cicero," and "Tully admired Tully," should be four ways of saying the same thing.[5]

If such a consequence of Mill's view is accepted, it would seem to have further consequences regarding 'intensional' contexts. Whether a sentence expresses a necessary truth or a contingent one depends only on the proposition expressed and not on the words used to express it. So any simple sentence should retain its 'modal value' (necessary, impossible, contingently true, or contingently false) when 'Cicero' is replaced by 'Tully' in one or more places, since such a replacement leaves the content of the sentence unaltered. Of course this implies that coreferential names are substitutable in modal contexts *salva veritate*: "It is necessary (possible) that Cicero . . ." and "It is necessary (possible) that Tully . . ." must have the same truth value no matter how the dots are filled by a simple sentence.

The situation would seem to be similar with respect to contexts involving knowledge, belief, and epistemic modalities. Whether a given subject believes something is presumably true or false of such a subject no matter how that belief is expressed; so if proper name substitution does not change the content of a sentence expressing a belief, coreferential proper names should be interchangeable *salva veritate* in belief contexts. Similar reasoning would hold for epistemic contexts ("Jones knows that . . .") and contexts of epistemic necessity ("Jones knows *a priori* that . . .") and the like.

All this, of course, would contrast strongly with the case of definite descriptions. It is well known that substitution of coreferential descriptions in simple sentences (without operators), on any reasonable conception of 'content,' *can* alter the content of such a sentence. In particular, the modal value of a sentence is not invariant under changes of coreferential descriptions: "The smallest prime is even" expresses a necessary truth, but "Jones's favorite number is even" expresses a contingent one, even if Jones's favorite number happens to be the smallest prime. It follows that coreferential descriptions are *not* interchangeable *salva veritate* in modal contexts: "It is necessary that the smallest prime is even" is true while "It is necessary that Jones's favorite number is even" is false.

Of course there is a '*de re*' or 'large scope' reading under which the second sentence is true. Such a reading would be expressed more accurately by

"Jones's favorite number is such that it is necessarily even" or, in rough Russellian transcription, as "One and only one number is admired by Jones above all others, and any such number is necessarily even (has the property of necessary evenness)." Such a *de re* reading, if it makes sense at all, by definition must be subject to a principle of substitution *salva veritate*, since necessary evenness is a property of the *number*, independently of how it is designated; in this respect there can be no contrast between names and descriptions. The contrast, according to the Millian view, must come in the *de dicto* or "small scope" reading, which is the *only* reading, for belief contexts as well as modal contexts, that will concern us in this paper. If we wish, we can emphasize that this is our reading in various ways. Say, "It is necessary that: Cicero was bald" or, more explicitly, "The following proposition is necessarily true: Cicero was bald," or even, in Carnap's 'formal' mode of speech,[6] " 'Cicero was bald' expresses a necessary truth." Now the Millian asserts that all these formulations retain their truth value when 'Cicero' is replaced by 'Tully,' even though 'Jones's favorite Latin author' and 'the man who denounced Catiline' would *not* similarly be interchangeable in these contexts even if they are codesignative.

Similarly for belief contexts. Here too *de re* beliefs — as in "Jones believes, *of* Cicero (or: *of* his favorite Latin author) that he was bald" do *not* concern us in this paper. Such contexts, if they make sense, are by definition subject to a substitutivity principle for both names and descriptions. Rather we are concerned with the *de dicto* locution expressed explicitly in such formulations as, "Jones believes that: Cicero was bald" (or: "Jones believes that: the man who denounced Catiline was bald"). The material after the colon expresses the *content* of Jones's belief. Other, more explicit, formulations are: "Jones believes the proposition — that — Cicero — was — bald," or even in the 'formal' mode, "The sentence 'Cicero was bald' gives the content of a belief of Jones." In all such contexts, the strict Millian seems to be committed to saying that codesignative names, but not codesignative descriptions, are interchangeable *salva veritate*.[7]

Now it has been widely assumed that these apparent consequences of the Millian view are plainly false. First, it seemed that sentences can alter their *modal* values by replacing a name by a codesignative one. "Hesperus is Hesperus" (or, more cautiously: "If Hesperus exists, Hesperus is Hesperus") expresses a necessary truth, while "Hesperus is Phosphorus" (or: "If Hesperus exists, Hesperus is Phosphorus"), expresses an empirical discovery, and hence, it has been widely assumed, a contingent truth. (It might have turned out, and hence might have been, otherwise.)

It has seemed even more obvious that codesignative proper names are not interchangeable in belief contexts and epistemic contexts. Tom, a normal speaker of the language, may sincerely assent to "Tully denounced Catiline," but not to "Cicero denounced Catiline." He may even deny the latter. And his denial is compatible with his status as a normal English speaker who satisfies normal criteria for using both 'Cicero' and 'Tully' as names for the famed Roman (without knowing that 'Cicero' and 'Tully' name the same person). Given this, it seems obvious that Tom believes that: Tully denounced Catiline, but that he does not believe (lacks the belief) that: Cicero denounced Catiline.[8] So it seems clear that codesignative proper names are not interchangeable in belief contexts. It also seems clear that there must be two distinct propositions or contents expressed by 'Cicero denounced Catiline' and 'Tully denounced Catiline.' How else can Tom believe one and deny the other? And the difference in propositions thus expressed can only come from a difference in *sense* between 'Tully' and 'Cicero.' Such a conclusion agrees with a Fregean theory and seems to be incompatible with a purely Millian view.[9]

In the previous work mentioned above, I rejected one of these arguments against Mill, the modal argument. 'Hesperus is Phosphorus,' I maintained, expresses just as necessary a truth as 'Hesperus is Hesperus'; there are no counterfactual situations in which Hesperus and Phosphorus would have been different. Admittedly, the truth of 'Hesperus is Phosphorus' was not known *a priori*, and may even have been widely disbelieved before appropriate empirical evidence came in. But these epistemic questions should be separated, I have argued, from the metaphysical question of the necessity of 'Hesperus is Phosphorus.' And it is a consequence of my conception of names as 'rigid designators' that codesignative proper names are interchangeable *salva veritate* in all contexts of (metaphysical) necessity and possibility; further, that replacement of a proper name by a codesignative name leaves the modal value of any sentence unchanged.

But although my position confirmed the Millian account of names in modal contexts, it equally appears at first blush to imply a *nonMillian* account of epistemic and belief contexts (and other contexts of propositional attitude). For I presupposed a sharp contrast between epistemic and metaphysical possibility: Before appropriate empirical discoveries were made, men might well have failed to know that Hesperus was Phosphorus, or even to believe it, even though they of course knew and believed that Hesperus was Hesperus.

Does not this support a Fregean position that 'Hesperus' and 'Phosphorus' have different 'modes of presentation' that determine their references? What else can account for the fact that, before astronomers identified the two heavenly bodies, a sentence using 'Hesperus' could express a common belief, while the same context involving 'Phosphorus' did not? In the case of 'Hesperus' and 'Phosphorus,' it is pretty clear what the different 'modes of presentation' would be: one mode determines a heavenly body by its typical position and appearance, in the appropriate season, in the evening; the other determines the same body by its position and appearance, in the appropriate season, in the morning. So it appears that even though, according to my view, proper names would be *modally* rigid — would have the same reference when we use them to speak of counterfactual situations as they do when used to describe the actual world — they would have a kind of Fregean 'sense' according to how that rigid reference is fixed. And the divergences of 'sense' (in this sense of 'sense') would lead to failures of interchangeability of co-designative names in contexts of propositional attitude, though not in modal contexts. Such a theory would agree with Mill regarding modal contexts but with Frege regarding belief contexts. The theory would not be *purely* Millian.[10]

After further thought, however, the Fregean conclusion appears less obvious. Just as people are said to have been unaware at one time of the fact that Hesperus is Phosphorus, so a normal speaker of English apparently may not know that Cicero is Tully, or that Holland is the Netherlands. For he may sincerely assent to 'Cicero was lazy,' while dissenting from 'Tully was lazy,' or he may sincerely assent to 'Holland is a beautiful country,' while dissenting from 'The Netherlands is a beautiful country.' In the case of 'Hesperus' and 'Phosphorus,' it seemed plausible to account for the parallel situation by supposing that 'Hesperus' and 'Phosphorus' fixed their (rigid) references to a single object in two conventionally different ways, one as the 'evening star' and one as the 'morning star.' But what corresponding *conventional* 'senses,' even taking 'senses' to be 'modes of fixing the reference rigidly,' can plausibly be supposed to exist for 'Cicero' and 'Tully' (or 'Holland' and 'the Netherlands')? Are not these just two names (in English) for the same man? Is there any special *conventional*, *community-wide* 'connotation' in the one lacking in the other?[11] I am unaware of any.[12]

Such considerations might seem to push us toward the extreme Frege-Russellian view that the senses of proper names vary, strictly speaking, from speaker to speaker, and that there is no community-wide sense but only a

community-wide reference.[13] According to such a view, the sense a given speaker attributes to such a name as 'Cicero' depends on which assertions beginning with 'Cicero' he accepts and which of these he regards as *defining*, for him, the name (as opposed to those he regards as mere factual beliefs 'about Cicero'). Similarly, for 'Tully.' For example, someone may define 'Cicero' as 'the Roman orator whose speech was Greek to Cassius,' and 'Tully' as 'the Roman orator who denounced Catiline.' Then such a speaker may well fail to accept 'Cicero is Tully' if he is unaware that a single orator satisfied both descriptions (if Shakespeare and history are both to be believed). He may well, in his ignorance, affirm 'Cicero was bald' while rejecting 'Tully was bald,' and the like. Is this not what actually occurs whenever someone's expressed beliefs fail to be indifferent to interchange of 'Tully' and 'Cicero'? Must not the source of such a failure lie in two distinct associated descriptions, or modes of determining the reference, of the two names? If a speaker does, as luck would have it, attach the same identifying properties both to 'Cicero' and to 'Tully,' he *will*, it would seem, use 'Cicero' and 'Tully' interchangeably. All this appears at first blush to be powerful support for the view of Frege and Russell that in general names are peculiar to idiolects, with 'senses' depending on the associated 'identifying descriptions.'

Note that, according to the view we are now entertaining, one *cannot* say, "Some people are unaware that Cicero is Tully.' For, according to this view, there is no single proposition denoted by the 'that' clause, that the community of normal English speakers expresses by 'Cicero is Tully.' Some — for example, those who define both 'Cicero' and 'Tully' as 'the author of *De Fato*' — use it to express a trivial self-identity. Others use it to express the proposition that the man who satisfied one description (say, that he denounced Catiline) is one and the same as the man who satisfied another (say, that his speech was Greek to Cassius). There is no single fact, 'that Cicero is Tully,' known by some but not all members of the community.

If I were to assert, "Many are unaware that Cicero is Tully," *I* would use 'that Cicero is Tully' to denote the proposition that *I* understand by these words. If this, for example, is a trivial self-identity, I would assert falsely, and irrelevantly, that there is widespread ignorance in the community of a certain self-identity.[14] I *can*, of course, say, "Some English speakers use both 'Cicero' and 'Tully' with the usual referent (the famed Roman) yet do not assent to 'Cicero is Tully.' "

This aspect of the Frege-Russellian view can, as before, be combined with a concession that names are rigid designators and that hence the description

used to fix the reference of a name is not synonymous with it. But there are considerable difficulties. There is the obvious intuitive unpalatability of the notion that we use such proper names as 'Cicero,' 'Venice,' 'Venus' (the planet) with differing 'senses' and for this reason do not 'strictly speaking' speak a single language. There are the many well-known and weighty objections to any description or cluster-of-descriptions theory of names. And is it definitely so clear that failure of interchangeability in belief contexts implies some difference of sense? After all, there is a considerable philosophical literature arguing that even word pairs that are straightforward synonyms if any pairs are — "doctor" and "physician," to give one example — are not interchangeable *salva veritate* in belief contexts, at least if the belief operators are iterated.[15]

A minor problem with this presentation of the argument for Frege and Russell will emerge in the next section: if Frege and Russell are right, it is not easy to state the very argument from belief contexts that appears to support them.

But the clearest objection, which shows that the others should be given their proper weight, is this: the view under consideration does not in fact account for the phenomena it seeks to explain. As I have said elsewhere,[16] individuals who "define 'Cicero' " by such phrases as "the Catiline denouncer," "the author of *De Fato*," etc., are relatively rare: their prevalence in the philosophical literature is the product of the excessive classical learning of some philosophers. Common men who clearly use 'Cicero' as a name for Cicero may be able to give no better answer to "Who was Cicero?" than "a famous Roman orator," and they probably would say the same (if anything!) for 'Tully.' (Actually, most people probably have never heard the name 'Tully.') Similarly, many people who have heard of both Feynman and Gell-Mann, would identify each as 'a leading contemporary theoretical physicist.' Such people do not assign 'senses' of the usual type to the names that uniquely identify the referent (even though they use the names with a determinate reference). But to the extent that the *indefinite* descriptions attached or associated can be called 'senses,' the 'senses' assigned to 'Cicero' and 'Tully,' or to 'Feynman' and 'Gell-Mann,' are *identical*.[17] Yet clearly speakers of this type can ask, "Were Cicero and Tully one Roman orator, or two different ones?" or "Are Feynman and Gell-Mann two different physicists, or one?" without knowing the answer to either question by inspecting 'senses' alone. Some such speaker might even conjecture, or be under the vague false impression, that, as he would say, 'Cicero was bald but Tully was not.' The

premise of the argument we are considering for the classic position of Frege and Russell — that whenever two codesignative names fail to be interchangeable in the expression of a speaker's beliefs, failure of interchangeability arises from a difference in the 'defining' descriptions the speaker associates with these names — is, therefore, false. The case illustrated by 'Cicero' and 'Tully' is, in fact, quite usual and ordinary. So the apparent failure of codesignative names to be everywhere interchangeable in belief contexts, is not to be explained by differences in the 'senses' of these names.

Since the extreme view of Frege and Russell does not in fact explain the apparent failure of the interchangeability of names in belief contexts, there seems to be no further reason — for present purposes — not to give the other overwhelming *prima facie* considerations against the Frege-Russell view their full weight. Names of famous cities, countries, persons, and planets are the common currency of our common language, not terms used homonymously in our separate idiolects.[18] The apparent failure of codesignative names to be interchangeable in belief contexts remains a mystery, but the mystery no longer seems so clearly to argue for a Fregean view as against a Millian one. Neither differing public senses nor differing private senses peculiar to each speaker account for the phenomena to be explained. So the apparent existence of such phenomena no longer gives a *prima facie* argument for such differing senses.

One final remark to close this section. I have referred before to my own earlier views in "Naming and Necessity." I said above that these views, inasmuch as they make proper names rigid and transparent[19] in modal contexts, favor Mill, but that the concession that proper names are not transparent in belief contexts appears to favor Frege. On a closer examination, however, the extent to which these opacity phenomena really support Frege against Mill becomes much more doubtful. And there are important theoretical reasons for viewing the "Naming and Necessity" approach in a Millian light. In that work I argued that ordinarily the real determinant of the reference of names of a former historical figure is a chain of communication, in which the reference of the name is passed from link to link. Now the legitimacy of such a chain accords much more with Millian views than with alternatives. For the view supposes that a learner acquires a name from the community by determining to use it with the same reference as does the community. We regard such a learner as using "Cicero is bald" to express the same thing the community expresses, regardless of variations in the properties different learners associate with 'Cicero,' as long as he determines that he will use the

name with the referent current in the community. That a name can be transmitted in this way accords nicely with a Millian picture, according to which only the reference, not more specific properties associated with the name, is relevant to the semantics of sentences containing it. It has been suggested that the chain of communication, which on the present picture determines the reference, might thereby itself be called a 'sense.' Perhaps so — if we wish[20] — but we should not thereby forget that the legitimacy of such a chain suggests that it is just preservation of reference, as Mill thought, that we regard as necessary for correct language learning.[21] (This contrasts with such terms as 'renate' and 'cordate,' where more than learning the correct extension is needed.) Also, as suggested above, the doctrine of rigidity in modal contexts is dissonant, though not necessarily inconsistent, with a view that invokes antiMillian considerations to explain propositional attitude contexts.

The spirit of my earlier views, then, suggests that a Millian line should be maintained as far as is feasible.

II. PRELIMINARIES: SOME GENERAL PRINCIPLES

Where are we now? We seem to be in something of a quandary. On the one hand, we concluded that the failure of 'Cicero' and 'Tully' to be interchangeable *salva veritate* in contexts of propositional attitude was by no means explicable in terms of different 'senses' of the two names. On the other hand, let us not forget the initial argument against Mill: If reference is *all there is* to naming, what semantic difference can there be between 'Cicero' and 'Tully'? And if there is no semantic difference, do not 'Cicero was bald' and 'Tully was bald' express exactly the same proposition? How, then, can anyone believe that Cicero was bald, yet doubt or disbelieve that Tully was?

Let us take stock. Why do we think that anyone can believe that Cicero was bald, but fail to believe that Tully was? Or believe, without any logical inconsistency, that Yale is a fine university, but that Old Eli is an inferior one? Well, a normal English speaker, Jones, can sincerely assent to 'Cicero was bald' but not to 'Tully was bald.' And this even though Jones uses 'Cicero' and 'Tully' in standard ways — he uses 'Cicero' in this assertion as a name for the Roman, not, say, for his dog, or for a German spy.

Let us make explicit the *disquotational principle* presupposed here, connecting sincere assent and belief. It can be stated as follows, where '*p*' is to be replaced, inside and outside all quotation marks, by any appropriate standard English sentence: "*If a normal English speaker, on reflection, sincerely assents to 'p,' then he believes that p.*" The sentence replacing '*p*' is to lack indexical or pronominal devices or ambiguities, that would ruin the intuitive sense of the principle (e.g., if he assents to "You are wonderful," he need not believe that *you* — the reader — are wonderful).[22] When we suppose that we are dealing with a normal speaker of English, we mean that he uses all words in the sentence in a standard way, combines them according to the appropriate syntax, etc.: in short, he uses the sentence to mean what a normal speaker should mean by it. The 'words' of the sentence may include proper names, where these are part of the common discourse of the community, so that we can speak of using them in a standard way. For example, if the sentence is "London is pretty," then the speaker should satisfy normal criteria for using 'London' as a name of London, and for using 'is pretty' to attribute an appropriate degree of pulchritude. The qualification "on reflection" guards against the possibility that a speaker may, through careless inattention to the meaning of his words or other momentary conceptual or linguistic confusion, assert something he does not really mean, or assent to a sentence in linguistic error. "Sincerely" is meant to exclude mendacity, acting, irony, and the like. I fear that even with all this it is possible that some astute reader — such, after all, is the way of philosophy — may discover a qualification I have overlooked, without which the asserted principle is subject to counterexample. I doubt, however, that any such modification will affect any of the uses of the principle to be considered below. Taken in its obvious intent, after all, the principle appears to be a self-evident truth. (A similar principle holds for sincere affirmation or assertion in place of assent.)

There is also a strengthened 'biconditional' form of the disquotational principle, where once again any appropriate English sentence may replace '*p*' throughout: *A normal English speaker who is not reticent will be disposed to sincere reflective assent to 'p' if and only if he believes that p.*[23] The biconditional form strengthens the simple one by adding that failure to assent indicates lack of belief, as assent indicates belief. The qualification about reticence is meant to take account of the fact that a speaker may fail to avow his beliefs because of shyness, a desire for secrecy, to avoid offense, etc. (An alternative formulation would give the speaker a sign to indicate lack of belief — not necessarily disbelief — in the assertion propounded, in addition to his sign of assent.) Maybe again the formulation needs further tightening, but the intent is clear.

Usually below the simple disquotational principle will be sufficient for our purposes, but once we will also invoke the strengthened form. The simple

form can often be used as a test for disbelief, provided the subject is a speaker with the modicum of logicality needed so that, at least after appropriate reflection, he does not hold simultaneously beliefs that are straightforward contradictions of each other — of the forms 'p' and '$\sim p$.'[24] (Nothing in such a requirement prevents him from holding simultaneous beliefs that jointly *entail* a contradiction.) In this case (where 'p' may be replaced by any appropriate English sentence), the speaker's assent to the negation of 'p' indicates not only his disbelief that p but also his failure to believe that p, using only the simple (unstrengthened) disquotational principle.

So far our principle applies only to speakers of English. It allows us to infer, from Peter's sincere reflective assent to "God exists," that he believes that God exists. But of course we ordinarily allow ourselves to draw conclusions, stated in English, about the beliefs of speakers of any language: we infer that Pierre believes that God exists from his sincere reflective assent to "*Dieu existe.*" There are several ways to do this, given conventional translations of French into English. We choose the following route. We have stated the disquotational principle in English, for English sentences; an analogous principle, stated in French (German, etc.) will be assumed to hold for French (German, etc.) sentences. Finally, we assume the *principle of translation*: *If a sentence of one language expresses a truth in that language, then any translation of it into any other language also expresses a truth* (*in that other language*). Some of our ordinary practice of translation may violate this principle; this happens when the translator's aim is not to preserve the content of the sentence, but to serve — in some other sense — the same purposes in the home language as the original utterance served in the foreign language.[25] But if the translation of a sentence is to mean the same as the sentence translated, preservation of truth value is a minimal condition that must be observed.

Granted the disquotational principle expressed in each language, reasoning starting from Pierre's assent to '*Dieu existe*' continues thus. First, on the basis of his utterance and the French disquotational principle we infer (in French):

Pierre croit que Dieu existe.

From this we deduce,[26] using the principle of translation:

Pierre believes that God exists.

In this way we can apply the disquotational technique to all languages.

Even if I apply the disquotational technique to English alone, there is a sense in which I can be regarded as tacitly invoking a principle of translation. For presumably I apply it to speakers of the language other than myself. As Quine has pointed out, to regard others as speaking the same language as I is in a sense tacitly to assume a *homophonic* translation of their language into my own. So when I infer from Peter's sincere assent to or affirmation of "God exists" that he believes that God exists, it is arguable that, strictly speaking, I combine the disquotational principle (for Peter's idiolect) with the principle of (homophonic) translation (of Peter's idiolect into mine). But for most purposes, we can formulate the disquotational principle for a single language, English, tacitly supposed to be the common language of English speakers. Only when the possibility of individual differences of dialect is relevant need we view the matter more elaborately.

Let us return from these abstractions to our main theme. Since a normal speaker — normal even in his use of 'Cicero' and 'Tully' as names — can give sincere and reflective assent to "Cicero was bald" and simultaneously to "Tully was not bald," the disquotational principle implies that he believes that Cicero was bald and believes that Tully was not bald. Since it seems that he need not have contradictory beliefs (even if he is a brilliant logician, he need not be able to deduce that at least one of his beliefs must be in error), and since a substitutivity principle for coreferential proper names in belief contexts would imply that he does have contradictory beliefs, it would seem that such a substitutivity principle must be incorrect. Indeed, the argument appears to be a *reductio ad absurdum* of the substitutivity principle in question.

The relation of this argument against substitutivity to the classical position of Russell and Frege is a curious one. As we have seen, the argument can be used to give *prima facie* support for the Frege-Russell view, and I think many philosophers have regarded it as such support. But in fact this very argument, which has been used to support Frege and Russell, cannot be stated in a straightforward fashion if Frege and Russell are right. For suppose Jones asserts, "Cicero was bald, but Tully was not." If Frege and Russell are right, I cannot deduce, using the disquotational principle:

(1) Jones believes that Cicero was bald but Tully was not,

since, in general, Jones and I will not, strictly speaking, share a common idiolect unless we assign the same 'senses' to all names. Nor can I combine disquotation and translation to the appropriate effect, since homophonic

translation of Jones's sentence into mine will in general be incorrect for the same reason. Since in fact I make no special distinction in sense between 'Cicero' and 'Tully' — to me, and probably to you as well, these are interchangeable names for the same man — and since according to Frege and Russell, Jones's very affirmation of (1) shows that for him there *is* some distinction of sense, Jones must therefore, on Frege-Russellian views, use one of these names differently from me, and homophonic translation is illegitimate. Hence, if Frege and Russell are right, we *cannot* use this example in the usual straightforward way to conclude that proper names are not substitutable in belief contexts — even though the example, and the ensuing negative verdict on substitutivity, has often been thought to support Frege and Russell!

Even according to the Frege-Russellian view, however, *Jones* can conclude, using the disquotational principle, and expressing his conclusion in his own idiolect:

(2) I believe that Cicero was bald but Tully was not.

I cannot endorse this conclusion in Jones's own words, since I do not share Jones's idiolect. I *can* of course conclude, "(2) expresses a truth in Jones's idiolect." I can also, if I find out the two 'senses' Jones assigns to 'Cicero' and 'Tully,' introduce two names 'X' and 'Y' into my own language with these same two senses ('Cicero' and 'Tully' have already been preempted) and conclude:

(3) Jones believes that X was bald and Y was not.

All this is enough so that we can still conclude, on the Frege-Russellian view, that codesignative names are not interchangeable in belief contexts. Indeed this can be shown more simply on this view, since codesignative descriptions plainly are not interchangeable in these contexts and for Frege and Russell names, being essentially abbreviated descriptions, cannot differ in this respect. Nevertheless, the simple argument, apparently free of such special Frege-Russellian doctrinal premises (and often used to support these premises), in fact cannot go through if Frege and Russell are right.

However, if, *pace* Frege and Russell, widely used names are common currency of our language, then there no longer is any problem for the simple argument, using the disquotational principle, to (2). So, it appears, on pain of convicting Jones of inconsistent beliefs — surely an unjust verdict — we must

not hold a substitutivity principle for names in belief contexts. If we used the *strengthened* disquotational principle, we could invoke Jones's presumed lack of any tendency to assent to 'Tully was bald' to conclude that he does not believe (lacks the belief) that Tully was bald. Now the refutation of the substitutivity principle is even stronger, for when applied to the conclusion that Jones believes that Cicero was bald but does not believe that Tully was bald, it would lead to a straightout contradiction. The contradiction would no longer be in Jones's beliefs but in our own.

This reasoning, I think, has been widely accepted as proof that codesignative proper names are not interchangeable in belief contexts. Usually the reasoning is left tacit, and it may well be thought that I have made heavy weather of an obvious conclusion. I wish, however, to question the reasoning. I shall do so without challenging any particular step of the argument. Rather I shall present — and this will form the core of the present paper — an argument for a paradox about names in belief contexts that invokes *no* principle of substitutivity. Instead it will be based on the principles — apparently so obvious that their use in these arguments is ordinarily tacit — of disquotation and translation.

Usually the argument will involve more than one language, so that the principle of translation and our conventional manual of translation must be invoked. We will also give an example, however, to show that a form of the paradox may result within English alone, so that the only principle invoked is that of disquotation (or, perhaps, disquotation plus *homophonic* translation). It will intuitively be fairly clear, in these cases, that the situation of the subject is 'essentially the same' as that of Jones with respect to 'Cicero' and 'Tully.' Moreover, the paradoxical conclusions about the subject will parallel those drawn about Jones on the basis of the substitutivity principle, and the arguments will parallel those regarding Jones. Only in these cases, no special substitutivity principle is invoked.

The usual use of Jones's case as a counterexample to the substitutivity principle is thus, I think, somewhat analogous to the following sort of procedure. Someone wishes to give a *reductio ad absurdum* argument against a hypothesis in topology. He does succeed in refuting this hypothesis, but his derivation of an absurdity from the hypothesis makes essential use of the unrestricted comprehension schema in set theory, which he regards as self-evident. (In particular, the class of all classes not members of themselves plays a key role in his argument.) Once we know that the unrestricted comprehension schema and the Russell class lead to contradiction by themselves, it is

clear that it was an error to blame the earlier contradiction on the topological hypothesis.

The situation would have been the same if, after deducing a contradiction from the topological hypothesis plus the 'obvious' unrestricted comprehension schema, it was found that a similar contradiction followed if we replaced the topological hypothesis by an apparently 'obvious' premise. In both cases it would be clear that, even though we may still not be confident of any specific flaw in the argument against the topological hypothesis, blaming the contradiction on that hypothesis is illegitimate: rather we are in a 'paradoxical' area where it is unclear *what* has gone wrong.[27]

It is my suggestion, then, that the situation with respect to the interchangeability of codesignative names is similar. True, such a principle, when combined with our normal disquotational judgments of belief, leads to straightforward absurdities. But we will see that the 'same' absurdities can be derived by replacing the interchangeability principle by our normal practices of translation and disquotation, or even by disquotation alone.

The particular principle stated here gives just one particular way of 'formalizing' our normal inferences from explicit affirmation or assent to belief; other ways of doing it are possible. It is undeniable that we *do* infer, from a normal Englishman's sincere affirmation of 'God exists' or 'London is pretty,' that he believes, respectively, that God exists or that London is pretty; and that we would make the same inferences from a Frenchman's affirmation of '*Dieu existe*' or '*Londres est jolie*.' Any principles that would justify such inferences are sufficient for the next section. It will be clear that the particular principles stated in the present section are sufficient, but in the next section the problem will be presented informally in terms of our inferences from foreign or domestic assertion to belief.

III. THE PUZZLE

Here, finally(!), is the puzzle. Suppose Pierre is a normal French speaker who lives in France and speaks not a word of English or of any other language except French. Of course he has heard of that famous distant city, London (which he of course calls '*Londres*') though he himself has never left France. On the basis of what he has heard of London, he is inclined to think that it is pretty. So he says, in French, "*Londres est jolie*."

On the basis of his sincere French utterance, we will conclude:

(4) Pierre believes that London is pretty.

I am supposing that Pierre satisfies all criteria for being a normal French speaker, in particular, that he satisfies whatever criteria we usually use to judge that a Frenchman (correctly) uses '*est jolie*' to attribute pulchritude and uses '*Londres*' — standardly — as a name of London.

Later, Pierre, through fortunate or unfortunate vicissitudes, moves to England, in fact to London itself, though to an unattractive part of the city with fairly uneducated inhabitants. He, like most of his neighbors, rarely ever leaves this part of the city. None of his neighbors know any French, so he must learn English by 'direct method,' without using any translation of English into French: by talking and mixing with the people he eventually begins to pick up English. In particular, everyone speaks of the city, 'London,' where they all live. Let us suppose for the moment — though we will see below that this is not crucial — that the local population are so uneducated that they know few of the facts that Pierre heard about London in France. Pierre learns from them everything they know about London, but there is little overlap with what he heard before. He learns, of course — speaking English — to call the city he lives in 'London.' Pierre's surroundings are, as I said, unattractive, and he is unimpressed with most of the rest of what he happens to see. So he is inclined to assent to the English sentence:

(5) London is not pretty.

He has *no* inclination to assent to:

(6) London is pretty.

Of course he does not for a moment withdraw his assent from the French sentence, "*Londres est jolie*"; he merely takes it for granted that the ugly city in which he is now stuck is distinct from the enchanting city he heard about in France. But he has no inclination to change his mind for a moment about the city he stills calls '*Londres*.'

This, then, is the puzzle. If we consider Pierre's past background as a French speaker, his entire linguistic behavior, on the same basis as we would draw such a conclusion about many of his countrymen, supports the conclusion ((4) above) that he believes that London is pretty. On the other hand, after Pierre lived in London for some time, he did not differ from his neighbors — his French background aside — either in his knowledge of English or in his command of the relevant facts of local geography. His English

vocabulary differs little from that of his neighbors. He, like them, rarely ventures from the dismal quarter of the city in which they all live. He, like them, knows that the city he lives in is called 'London' and knows a few other facts. Now Pierre's neighbors would surely be said to use 'London' as a name for London and to speak English. Since, as an English speaker, he does not differ at all from them, we should say the same of him. But then, on the basis of his sincere assent to (5), we should conclude:

(7) Pierre believes that London is not pretty.

How can we describe this situation? It seems undeniable that Pierre *once* believed that London is pretty — at least before he learned English. For at that time, he differed not at all from countless numbers of his countrymen, and we would have exactly the same grounds to say of him as of any of them that he believes that London is pretty: if any Frenchman who was both ignorant of English and never visited London believed that London is pretty, Pierre did. Nor does it have any plausibility to suppose, because of his later situation *after* he learns English, that Pierre should *retroactively* be judged *never* to have believed that London is pretty. To allow such *ex post facto* legislation would, as long as the future is uncertain, endanger our attributions of belief to *all* monolingual Frenchmen. We would be forced to say that Marie, a monolingual who firmly and sincerely asserts, "*Londres est jolie*," may or may not believe that London is pretty depending on the *later* vicissitudes of her career (if later she learns English and . . ., . . .). No: Pierre, like Marie, believed that London is pretty when he was monolingual.

Should we say that Pierre, now that he lives in London and speaks English, no longer believes that London is pretty? Well, unquestionably Pierre *once* believed that London is pretty. So we would be forced to say that Pierre has *changed his mind, has given up his previous belief*. But has he really done so? Pierre is very set in his ways. He reiterates, with vigor, every assertion he has ever made in French. He says he has not changed his mind about anything, has *not* given up any belief. Can we say he is wrong about this? If we did not have the story of his living in London and his English utterances, on the basis of his normal command of French we would be *forced* to conclude that he *still* believes that London is pretty. And it does seem that this is correct. Pierre has neither changed his mind nor given up any belief he had in France.

Similar difficulties beset any attempt to deny him his new belief. His French past aside, he is just like his friends in London. Anyone else, growing up in London with the same knowledge and beliefs that he expresses in England, we would undoubtedly judge to believe that London is not pretty. Can Pierre's French past nullify such a judgment? Can we say that Pierre, because of his French past, does not believe that (5)? Suppose an electric shock wiped out all his memories of the French language, what he learned in France, and his French past. He would then be *exactly* like his neighbors in London. He would have the *same* knowledge, beliefs, and linguistic capacities. We then presumably would be forced to say that Pierre believes that London is ugly if we say it of his neighbors. But surely no shock that *destroys* part of Pierre's memories and knowledge can *give* him a new belief. If Pierre believes (5) *after* the shock, he believed it before, despite his French language and background.

If we would deny Pierre, in his bilingual stage, his belief that London is pretty *and* his belief that London is not pretty, we combine the difficulties of both previous options. We still would be forced to judge that Pierre once believed that London is pretty but does no longer, in spite of Pierre's own sincere denial that he has lost any belief. We also must worry whether Pierre would *gain* the belief that London is not pretty if he totally forgot his French past. The option does not seem very satisfactory.

So now it seems that we must respect both Pierre's French utterances and their English counterparts. So we must say that Pierre has contradictory beliefs, that he believes that London is pretty *and* he believes that London is not pretty. But there seem to be insuperable difficulties with this alternative as well. We may suppose that Pierre, in spite of the unfortunate situation in which he now finds himself, is a leading philosopher and logician. He would *never* let contradictory beliefs pass. And surely anyone, leading logician or no, is in principle in a position to notice and correct contradictory beliefs if he has them. Precisely for this reason, we regard individuals who contradict themselves as subject to greater censure than those who merely have false beliefs. But it is clear that Pierre, as long as he is unaware that the cities he calls 'London' and '*Londres*' are one and the same, is in no position to see, by logic alone, that at least one of his beliefs must be false. He lacks information, not logical acumen. He cannot be convicted of inconsistency: to do so is incorrect.

We can shed more light on this if we change the case. Suppose that, in France, Pierre, instead of affirming "*Londres est jolie*," had affirmed, more cautiously, "*Si New York est jolie, Londres est jolie aussi*," so that he believed that *if* New York is pretty, so is London. Later Pierre moves to London,

learns English as before, and says (in English) "London is not pretty." So he now believes, further, that London is *not* pretty. Now from the two premises, both of which appears to be among his beliefs (a) If New York is pretty, London is, and (b) London is not pretty, Pierre should be able to deduce by *modus tollens* that New York is not pretty. But no matter how great Pierre's logical acumen may be, *he cannot in fact make any such deduction, as long as he supposes that 'Londres' and 'London' may name two different cities.* If he *did* draw such a conclusion, he would be guilty of a fallacy.

Intuitively, he may well suspect that New York is pretty, and just this suspicion may lead him to suppose that '*Londres*' and 'London' probably name distinct cities. Yet, if we follow our normal practice of reporting the beliefs of French and English speakers, *Pierre has available to him (among his beliefs) both the premises of a modus tollens argument that New York is not pretty.*

Again, we may emphasize Pierre's *lack* of belief instead of his belief. Pierre, as I said, has no disposition to assent to (6). Let us concentrate on this, ignoring his disposition to assent to (5). In fact, if we wish we may change the case: Suppose Pierre's neighbors think that since they rarely venture outside their own ugly section, they have no right to any opinion as to the pulchritude of the whole city. Suppose Pierre shares their attitude. Then, judging by his failure to respond affirmatively to "London is pretty," we may judge, from Pierre's behavior as an *English* speaker, that he lacks the belief that London is pretty: never mind whether he disbelieves it, as before, or whether, as in the modified story, he insists that he has no firm opinion on the matter.

Now (using the *strengthened* disquotational principle), we can derive a contradiction, not merely in Pierre's judgments, but in our own. For on the basis of his behavior as an English speaker, we concluded that he does *not* believe that London is pretty (that is, that it is not the case that he believes that London is pretty). But on the basis of his behavior as a *French* speaker, we must conclude that he *does* believe that London is pretty. This is a contradiction.[28]

We have examined four possibilities for characterizing Pierre while he is in London: (a) that at that time we no longer respect his French utterance ('*Londres est jolie*'), that is that we no longer ascribe to him the corresponding belief; (b) that we do not respect his English utterance (or lack of utterance); (c) that we respect neither; (d) that we respect both. Each possibility seems to lead us to say something either plainly false or even downright contradictory. Yet the possibilities appear to be logically exhaustive. This, then, is the paradox.

I have no firm belief as to how to solve it. But beware of one source of confusion. It is no solution in itself to observe that some *other* terminology, which evades the question whether Pierre believes that London is pretty, may be sufficient to state all the relevant facts. I am fully aware that complete and straightforward descriptions of the situation are possible and that in this sense there is no paradox. Pierre is disposed to sincere assent to '*Londres est jolie*' but not to 'London is pretty.' He uses French normally, English normally. Both with '*Londres*' and 'London' he associates properties sufficient to determine that famous city, but he does not realize that they determine a single city. (And his uses of '*Londres*' and 'London' are historically (causally) connected with the same single city, though he is unaware of that.) We may even give a rough statement of his beliefs. He believes that the city he calls '*Londres*' is pretty, that the city he calls 'London' is not. No doubt other straightforward descriptions are possible. No doubt some of these are, in a certain sense, *complete* descriptions of the situation.

But none of this answers the original question. Does Pierre, or does he not, believe that London is pretty? I know of no answer to *this* question that seems satisfactory. It is no answer to protest that, in some *other* terminology, one can state 'all the relevant facts.'

To reiterate, this is the puzzle: Does Pierre, or does he not, believe that London is pretty? It is clear that our normal criteria for the attribution of belief lead, when applied to *this* question, to paradoxes and contradictions. One set of principles adequate to many ordinary attributions of belief, but which leads to paradox in the present case, was stated in Section 2; and other formulations are possible. As in the case of the logical paradoxes, the present puzzle presents us with a problem for customarily accepted principles and a challenge to formulate an acceptable set of principles that does not lead to paradox, is intuitively sound, and supports the inferences we usually make. Such a challenge cannot be met simply by a description of Pierre's situation that evades the question whether he believes that London is pretty.

One aspect of the presentation may misleadingly suggest the applicability of Frege-Russellian ideas that each speaker associates his own description or properties to each name. For as I just set up the case Pierre learned one set of facts about the so-called '*Londres*' when he was in France, and *another* set of facts about 'London' in England. Thus it may appear that 'what's really going on' is that Pierre believes that *the city* satisfying *one* set of properties *is*

pretty, while he believes that *the city* satisfying *another* set of properties *is not* pretty.

As we just emphasized, the phrase 'what's really going on' is a danger signal in discussions of the present paradox. The conditions stated may — let us concede for the moment — describe 'what's really going on.' But they do not resolve the problem with which we began, that of the behavior of names in belief contexts: Does Pierre, or does he not, believe that London (not the city satisfying such-and-such descriptions, but *London*) is pretty? No answer has yet been given.

Nevertheless, these considerations may appear to indicate that descriptions, or associated properties, are highly relevant somehow to an ultimate solution, since at this stage it appears that the entire puzzle arises from the fact that Pierre originally associated different identifying properties with 'London' and '*Londres*.' Such a reaction may have some force even in the face of the now fairly well-known arguments against 'identifying descriptions' as in any way 'defining,' or even 'fixing the reference' of names. But in fact the special features of the case, as I set it out, are misleading. The puzzle can arise even if Pierre associates exactly the same identifying properties with both names.

First, the considerations mentioned above in connection with 'Cicero' and 'Tully' establish this fact. For example, Pierre may well learn, in France, '*Platon*' as the name of a major Greek philosopher, and later, in England, learns 'Plato' with the same identification. Then the same puzzle can arise: Pierre may have believed, when he was in France and was monolingual in French, that Plato was bald (he would have said, "*Platon était chauve*"), and later conjecture, in English, "Plato was not bald," thus indicating that he believes or suspects that Plato was *not* bald. He need only suppose that, in spite of the similarity of their names, the man he calls '*Platon*' and the man he calls 'Plato' were two distinct major Greek philosophers. In principle, the same thing could happen with 'London' and '*Londres*.'

Of course, most of us learn a *definite* description about London, say 'the largest city in England.' Can the puzzle still arise? It is noteworthy that the puzzle can still arise even if Pierre associates to '*Londres*' and to 'London' *exactly* the same *uniquely identifying* properties. How can this be? Well, suppose that Pierre believes that London is the largest city in (and capital of) England, that it contains Buckingham Palace, the residence of the Queen of England, and he believes (correctly) that these properties, conjointly, uniquely identify the city. (In this case, it is best to suppose that he has never

seen London, or even England, so that he uses *only* these properties to identify the city. Nevertheless, he has learned English by 'direct method.') These uniquely identifying properties he comes to associate with 'London' after he learned English, and he expresses the appropriate beliefs about 'London' in English. Earlier, when he spoke nothing but French, however, he associated *exactly* the same uniquely identifying properties with '*Londres*.' He believed that '*Londres*,' as he called it, could be uniquely identified as the capital of England, that it contained Buckingham Palace, that the Queen of England lived there, etc. Of course he expressed these beliefs, like most monolingual Frenchmen, in French. In particular, he used '*Angleterre*' for England, '*le Palais de Buckingham*' (pronounced '*Bookeengam*'!) for Buckingham Palace, and '*la Reine d'Angleterre*' for the Queen of England. But if any Frenchman who speaks no English can ever be said to associate *exactly* the properties of being the capital of England etc., with the name '*Londres*,' Pierre in his monolingual period did so.

When Pierre becomes a bilingual, *must* he conclude that 'London' and '*Londres*' name the same city, because he defined each by the same uniquely identifying properties?

Surprisingly, no! Suppose Pierre had affirmed, '*Londres est jolie*.' If Pierre has any reason — even just a 'feeling in his bones,' or perhaps exposure to a photograph of a miserable area which he was told (in English) was part of 'London' — to maintain 'London is not pretty,' he need not contradict himself. He need only conclude that 'England' and '*Angleterre*' name two different countries, that 'Buckingham Palace' and '*le Palais de Buckingham*' (recall the pronunciation!), name two different palaces, and so on. Then he can maintain *both* views without contradiction, and regard *both* properties as uniquely identifying.

The fact is that the paradox reproduces itself on the level of the 'uniquely identifying properties' that description theorists have regarded as 'defining' proper names (and *a fortiori*, as fixing their references). Nothing is more reasonable than to suppose that if two names, A and B, and a single set of properties, S, are such that a certain speaker believes that the referent of A uniquely satisfies all of S and that the referent of B also uniquely satisfies all of S, then that speaker is committed to the belief that A and B have the same reference. In fact, the identity of the referents of A and B is an easy *logical consequence* of the speaker's beliefs.

From this fact description theorists concluded that names can be regarded as synonymous, and hence interchangeable *salva veritate* even in belief con-

texts, provided that they are 'defined' by the same uniquely identifying properties.

We have already seen that there is a difficulty in that the set S of properties need not in fact be uniquely identifying. But in the present paradoxical situation there is a surprising difficulty even if the supposition of the description theorist (that the speaker believes that S is uniquely fulfilled) in fact holds. For, as we have seen above, Pierre is in no position to draw ordinary logical consequences from the conjoint set of what, when we consider him separately as a speaker of English and as a speaker of French, we would call his beliefs. He cannot infer a contradiction from his separate beliefs that London is pretty and that London is not pretty. Nor, in the modified situation above, would Pierre make a normal *modus tollens* inference from his beliefs that London is not pretty and that London is pretty if New York is. Similarly here, if we pay attention only to Pierre's behavior as a French speaker (and at least in his monolingual days he was no different from any other Frenchmen), Pierre satisfies all the normal criteria for believing that '*Londres*' has a referent uniquely satisfying the properties of being the largest city in England, containing Buckingham Palace, and the like. (If Pierre did not hold such beliefs, no Frenchman *ever* did.) Similarly, on the basis of his (later) beliefs expressed in English, Pierre also believes that the referent of 'London' uniquely satisfies these same properties. But Pierre cannot combine the two beliefs into a single set of beliefs from which he can draw the normal conclusion that 'London' and '*Londres*' must have the same referent. (Here the trouble comes not from 'London' and '*Londres*' but from 'England' and '*Angleterre*' and the rest.) Indeed, if he *did* draw what would appear to be the normal conclusion in this case and any of the other cases, Pierre would in fact be guilty of a logical fallacy.

Of course the description theorist could hope to eliminate the problem by 'defining' '*Angleterre*,' 'England,' and so on by appropriate descriptions also. Since in principle the problem may rear its head at the next 'level' and at each subsequent level, the description theorist would have to believe that an 'ultimate' level can eventually be reached where the defining properties are 'pure' properties not involving proper names (nor natural kind terms or related terms, see below!). I know of no convincing reason to suppose that such a level can be reached in any plausible way, or that the properties can continue to be uniquely identifying if one attempts to eliminate all names and related devices.[29] Such speculation aside, the fact remains that Pierre, judged by the *ordinary* criteria for such judgments, *did* learn both '*Londres*' and

'London' by *exactly* the same set of identifying properties; yet the puzzle remains even in this case.

Well, then, is there any way out of the puzzle? Aside from the principles of disquotation and translation, only our normal practice of translation of French into English has been used. Since the principles of disquotation and translation seem self-evident, we may be tempted to blame the trouble on the translation of '*Londres est jolie*' as 'London is pretty,' and ultimately, then, on the translation of '*Londres*' as 'London.'[30] Should we, perhaps, permit ourselves to conclude that '*Londres*' should not, 'strictly speaking' be translated as 'London'? Such an expedient is, of course, desperate: the translation in question is a standard one, learned by students together with other standard translations of French into English. Indeed, '*Londres*' is, in effect, introduced into French as the French version of 'London.'

Since our backs, however, are against the wall, let us consider this desperate and implausible expedient a bit further. If '*Londres*' is *not* a correct French version of the English 'London,' under what circumstances can proper names be translated from one language to another?

Classical description theories suggest the answer: Translation, strictly speaking, is between idiolects; a name in one idiolect can be translated into another when (and only when) the speakers of the two idiolects associate the same uniquely identifying properties with the two names. We have seen that any such proposed restriction, not only fails blatantly to fit our normal practices of translation and indirect discourse reportage, but does not even appear to block the paradox.[31]

So we still want a suitable restriction. Let us drop the references to idiolects and return to '*Londres*' and 'London' as names in French and English, respectively — the languages of two communities. If '*Londres*' is not a correct French translation of 'London,' could any other version do better? Suppose I introduced another word into French, with the stipulation that *it* should always be used to translate 'London.' Would not the same problem arise for this word as well? The only feasible solution in this direction is the most drastic: decree that no sentence containing a name can be translated except by a sentence containing the phonetically identical name. Thus when Pierre asserts '*Londres est jolie*,' we English speakers can at best conclude, if anything: Pierre believes that *Londres* is pretty. Such a conclusion is, of course, not expressed in English, but in a word salad of English and French; on the view now being entertained, we cannot state Pierre's belief in *English* at all.[32] Similarly, we would have to say: Pierre believes that *Angleterre* is a

monarchy, Pierre believes that *Platon* wrote dialogues, and the like.[33]

This 'solution' appears at first to be effective against the paradox, but it is drastic. What is it about sentences containing names that makes them — a substantial class — intrinsically untranslatable, express beliefs that cannot be reported in any other language? At best, to report them in the other language, one is forced to use a word salad in which names from the one language are imported into the other. Such a supposition is both contrary to our normal practice of translation and very implausible on its face.

Implausible though it is, there is at least this much excuse for the 'solution' at this point. Our normal practice with respect to some famous people and especially for geographical localities is to have different names for them in different languages, so that in translating sentences we translate the names. But for a large number of names, especially names of people, this is not so: the person's name is used in the sentences of all languages. At least the restriction in question merely urges us to mend our ways by doing *always* what we presently do *sometimes*.

But the really drastic character of the proposed restriction comes out when we see how far it may have to extend. In "Naming and Necessity" I suggested that there are important analogies between proper names and natural kind terms, and it seems to me that the present puzzle is one instance where the analogy will hold. Putnam, who has proposed views on natural kinds similar to my own in many respects, stressed this extension of the puzzle in his comments at the Conference. Not that the puzzle extends to all translations from English to French. At the moment, at least, it seems to me that Pierre, if he learns English and French separately, without learning any translation manual between them, *must* conclude, if he reflects enough, that 'doctor' and '*médecin*,' and '*heureux*' and 'happy,' are synonymous, or at any rate, coextensive;[34] any potential paradox of the present kind for these word pairs is thus blocked. But what about '*lapin*' and 'rabbit,' or 'beech' and '*hêtre*'? We may suppose that Pierre is himself neither a zoologist nor a botanist. He has learned each language in its own country and the examples he has been shown to illustrate '*les lapins*' and 'rabbits,' 'beeches' and '*les hêtres*' are distinct. It thus seems to be possible for him to suppose that '*lapin*' and 'rabbit,' or 'beech' and '*hêtre*,' denote distinct but superficially similar kinds or species, even though the differences may be indiscernible to the untrained eye. (This is especially plausible if, as Putnam supposes, an English speaker — for example, Putnam himself — who is not a botanist may use 'beech' and 'elm' with their normal (distinct) meanings, even though he cannot himself

distinguish the two trees.[35] Pierre may quite plausibly be supposed to wonder whether the trees which in France he called '*les hêtres*' were beeches or elms, even though as a speaker of French he satisfies all usual criteria for using '*les hêtres*' normally. If beeches and elms will not serve, better pairs of ringers exist that cannot be told apart except by an expert.) Once Pierre is in such a situation, paradoxes analogous to the one about London obviously can arise for rabbits and beeches. Pierre could affirm a French statement with '*lapin*,' but deny its English translation with 'rabbit.' As above, we are hard-pressed to say what Pierre *believes*. We were considering a 'strict and philosophical' reform of translation procedures which proposed that foreign proper names should always be appropriated rather than translated. Now it seems that we will be forced to do the same with all words for natural kinds. (For example, on price of paradox, one must not translate '*lapin*' as 'rabbit'!) No longer can the extended proposal be defended, even weakly, as 'merely' universalizing what we already do sometimes. It is surely too drastic a change to retain any credibility.[36]

There is yet another consideration that makes the proposed restriction more implausible: Even this restriction does not really block the paradox. Even if we confine ourselves to a single language, say English, and to phonetically identical tokens of a single name, we can still generate the puzzle. Peter (as we may as well say now) may learn the name 'Paderewski' with an identification of the person named as a famous pianist. Naturally, having learned this, Peter will assent to "Paderewski had musical talent," and *we* can infer — using 'Paderewski,' as we usually do, to name the Polish musician and statesman:

(8) Peter believes that Paderewski had musical talent.

Only the disquotational principle is necessary for our inference; no translation is required. Later, in a different circle, Peter learns of someone called 'Paderewski' who was a Polish nationalist leader and Prime Minister. Peter is skeptical of the musical abilities of politicians. He concludes that probably two people, approximate contemporaries no doubt, were both named 'Paderewski.' Using 'Paderewski' as a name for the *statesman*, Peter assents to, "Paderewski had no musical talent." Should we infer, by the disquotational principle,

(9) Peter believes that Paderewski had no musical talent

or should we not? If Peter had not had the past history of learning the name

'Paderewski' in another way, we certainly would judge him to be using 'Paderewski' in a normal way, with the normal reference, and we would infer (9) by the disquotational principle. The situation is parallel to the problem with Pierre and London. Here, however, no restriction that names should not be translated, but should be phonetically repeated in the translation, can help us. Only a single language and a single name are involved. If any notion of translation is involved in this example, it is homophonic translation. Only the disquotational principle is used explicitly.[37] (On the other hand, the original 'two languages' case had the advantage that it would apply even if we spoke languages in which all names must denote uniquely and unambiguously.) The restriction that names must not be translated is thus ineffective, as well as implausible and drastic.

I close this section with some remarks on the relation of the present puzzle to Quine's doctrine of the 'indeterminacy of translation,' with its attendant repudiation of intensional idioms of 'propositional attitude' such as belief and even indirect quotation. To a sympathizer with these doctrines the present puzzle may well seem to be just more grist for a familiar mill. The situation of the puzzle seems to lead to a breakdown of our normal practices of attributing belief and even of indirect quotation. No obvious paradox arises if we describe the same situation in terms of Pierre's sincere assent to various sentences, together with the conditions under which he has learned the name in question. Such a description, although it does not yet conform to Quine's strict behavioristic standards, fits in well with his view that in some sense direct quotation is a more 'objective' idiom than the propositional attitudes. Even those who, like the present writer, do not find Quine's negative attitude to the attitudes completely attractive must surely acknowledge this.

But although sympathizers with Quine's view can use the present examples to support it, the differences between these examples and the considerations Quine adduces for his own skepticism about belief and translation should not escape us. Here we make no use of hypothetical exotic systems of translation differing radically from the usual one, translating '*lapin*,' say, as 'rabbit stage' or 'undetached part of a rabbit.' The problem arises entirely within our usual and customary sytem of translation of French into English; in one case, the puzzle arose even within English alone, using at most 'homophonic' translation. Nor is the problem that many different interpretations or translations fit our usual criteria, that, in Davidson's phrase,[38] there is more than one 'way of getting it right.' The trouble here is not that many views as to Pierre's beliefs get it right, but that they all definitely get it *wrong*. A straightforward

application of the principles of translation and disquotation to all Pierre's utterances, French and English, yields the result that Pierre holds inconsistent beliefs, that logic alone should teach him that one of his beliefs is false. Intuitively, this is plainly incorrect. If we refuse to apply the principles to his French utterances at all, we would conclude that Pierre never believed that London is pretty, even though, before his unpredictable move, he was like any other monolingual Frenchman. This is absurd. If we refuse to ascribe the belief in London's pulchritude only after Pierre's move to England, we get the counterintuitive result that Pierre has changed his mind, and so on. But we have surveyed the possibilities above: the point was not that they are 'equally good,' but that all are *obviously wrong*. If the puzzle is to be used as an argument for a Quinean position, it is an argument of a fundamentally different kind from those given before. And even Quine, if he wishes to incorporate the notion of belief even into a 'second level' of canonical notation,[39] must regard the puzzle as a real problem.

The alleged indeterminacy of translation and indirect quotation causes relatively little trouble for such a scheme for belief; the embarrassment it presents to such a scheme is, after all, one of riches. But the present puzzle indicates that the usual principles we use to ascribe beliefs are apt, in certain cases, to lead to contradiction, or at least, patent falsehoods. So it presents a problem for any project, Quinean or other, that wishes to deal with the 'logic' of belief on any level.[40]

IV. CONCLUSION

What morals can be drawn? The primary moral — quite independent of any of the discussion of the first two sections — is that the puzzle *is* a puzzle. As any theory of truth must deal with the Liar Paradox, so any theory of belief and names must deal with this puzzle.

But our theoretical starting point in the first two sections concerned proper names and belief. Let us return to Jones, who assents to "Cicero was bald" and to "Tully was not bald." Philosophers, using the disquotational principle, have concluded that Jones believes that Cicero was bald but that Tully was not. Hence, they have concluded, since Jones does not have contradictory beliefs, belief contexts are not 'Shakespearean' in Geach's sense: codesignative proper names are not interchangeable in these contexts *salva veritate*.[41]

I think the puzzle about Pierre shows that the simple conclusion was unwarranted. Jones' situation strikingly resembles Pierre's. A proposal that

'Cicero' and 'Tully' *are* interchangeable amounts roughly to a homophonic 'translation' of English into itself in which 'Cicero' is mapped into 'Tully' and *vice versa*, while the rest is left fixed. Such a 'translation' can, indeed, be used to obtain a paradox. But should the problem be blamed on this step? Ordinarily we would suppose without question that sentences in French with '*Londres*' should be translated into English with 'London.' Yet the same paradox results when we apply this translation too. We have seen that the problem can even arise with a single name in a single language, and that it arises with natural kind terms in two languages (or one: see below).

Intuitively, Jones' assent to both 'Cicero was bald' and 'Tully was not bald' arises from sources of just the same kind as Pierre's assent to both '*Londres est jolie*' and 'London is not pretty.'

It is wrong to blame unpalatable conclusions about Jones on substitutivity. The reason does not lie in any specific fallacy in the argument but rather in the nature of the realm being entered. Jones's case is just like Pierre's: both are in an area where our normal practices of attributing belief, based on the principles of disquotation and translation or on similar principles, are questionable.

It should be noted in this connection that the principles of disquotation and translation can lead to 'proofs' as well as 'disproofs' of substitutivity in belief contexts. In Hebrew there are two names for Germany, transliteratable roughly as '*Ashkenaz*' and '*Germaniah*' — the first of these may be somewhat archaic. When Hebrew sentences are translated into English, both become 'Germany.' Plainly a normal Hebrew speaker analogous to Jones might assent to a Hebrew sentence involving '*Ashkenaz*' while dissenting from its counterpart with '*Germaniah.*' So far there is an argument *against* substitutivity. But there is also an argument *for* substitutivity, based on the principle of translation. Translate a Hebrew sentence involving '*Ashkenaz*' into English, so that '*Ashkenaz*' goes into 'Germany.' Then retranslate the result into Hebrew, this time translating 'Germany' as '*Germaniah.*' By the principle of translation, both translations preserve truth value. So: the truth value of any sentence of Hebrew involving '*Ashkenaz*' remains the same when '*Ashkenaz*' is replaced by '*Germaniah*' — a 'proof' of substitutivity! A similar 'proof' can be provided wherever there are two names in one language, and a normal practice of translating both indifferently into a single name of another language.[42] (If we combine the 'proof' and 'disproof' of substitutivity in this paragraph, we could get yet another paradox analogous to Pierre's: our Hebrew speaker both believes, and disbelieves, that Germany is pretty. Yet

no amount of pure logic or semantic introspection suffices for him to discover his error.)

Another consideration, regarding natural kinds: Previously we pointed out that a bilingual may learn '*lapin*' and 'rabbit' normally in each respective language yet wonder whether they are one species or two, and that this fact can be used to generate a paradox analogous to Pierre's. Similarly, a speaker of *English* alone may learn 'furze' and 'gorse' normally (separately), yet wonder whether these are the same, or resembling kinds. (What about 'rabbit' and 'hare'?) It would be easy for such a speaker to assent to an assertion formulated with 'furze' but withhold assent from the corresponding assertion involving 'gorse.' The situation is quite analogous to that of Jones with respect to 'Cicero' and 'Tully.' Yet 'furze' and 'gorse,' and other pairs of terms for the same natural kind, are normally thought of as *synonyms*.

The point is *not*, of course, that codesignative proper names *are* interchangeable in belief contexts *salva veritate*, or that they *are* interchangeable in simple contexts even *salva significatione*. The point is that the absurdities that disquotation plus substitutivity would generate are exactly paralleled by absurdities generated by disquotation plus translation, or even 'disquotation alone' (or: disquotation plus homophonic translation). Also, though our naive practice may lead to 'disproofs' of substitutivity in certain cases, it can also lead to 'proofs' of substitutivity in some of these same cases, as we saw two paragraphs back. When we enter into the area exemplified by Jones and Pierre, we enter into an area where our normal practices of interpretation and attribution of belief are subjected to the greatest possible strain, perhaps to the point of breakdown. So is the notion of the *content* of someone's assertion, the *proposition* it expresses. In the present state of our knowledge, I think it would be foolish to draw any conclusion, positive or negative, about substitutivity.[43]

Of course nothing in these considerations prevents us from observing that Jones can sincerely assert both "Cicero is bald" and "Tully is not bald," even though he is a normal speaker of English and uses 'Cicero' and 'Tully' in normal ways, and with the normal referent. Pierre and the other paradoxical cases can be described similarly. (For those interested in one of my own doctrines, we can still say that there was a time when men were in no epistemic position to assent to 'Hesperus is Phosphorus' for want of empirical information, but it nevertheless expressed a necessary truth.)[44] But it is no surprise that quoted contexts fail to satisfy a substitutivity principle within the quotation marks. And, in our *present* state of clarity about the problem, we are in

no position to apply a disquotation principle to these cases, nor to judge when two such sentences do, or do not, express the same 'proposition.'

Nothing in the discussion impugns the conventional judgment that belief contexts are 'referentially opaque,' if 'referential opacity' is construed so that failure of coreferential *definite descriptions* to be interchangeable *salva veritate* is sufficient for referential opacity. No doubt Jones can believe that the number of planets is even, without believing that the square of three is even, if he is under a misapprehension about the astronomical, but not the arithmetical facts. The question at hand was whether belief contexts were 'Shakespearean,' not whether they were 'referentially transparent.' (Modal contexts, in my opinion, are 'Shakespearean' but 'referentially opaque.')[45]

Even were we inclined to rule that belief contexts are not Shakespearean, it would be implausible at present to use the phenomenon to support a Frege-Russellian theory that names have descriptive 'senses' through 'uniquely identifying properties.' There are the well-known arguments against description theories, independent of the present discussion; there is the implausibility of the view that difference in names is difference in idiolect; and finally, there are the arguments of the present paper that differences of associated properties do not explain the problems in any case. Given these considerations, and the cloud our paradox places over the notion of 'content' in this area, the relation of substitutivity to the dispute between Millian and Fregean conclusions is not very clear.

We repeat our conclusions: Philosophers have often, basing themselves on Jones' and similar cases, supposed that it goes virtually without saying that belief contexts are not 'Shakespearean.' I think that, at present, such a definite conclusion is unwarranted. Rather Jones' case, like Pierre's, lies in an area where our normal apparatus for the ascription of belief is placed under the greatest strain and may even break down. There is even less warrant at the present time, in the absence of a better understanding of the paradoxes of this paper, for the use of alleged failures of substitutivity in belief contexts to draw any significant theoretical conclusion about proper names. Hard cases make bad law.[46]

Princeton University

NOTES

1 "Naming and Necessity," in: *The Semantics of Natural Languages*, D. Davidson and G. Harman (eds.), Dordrecht, Reidel, 1971, pp. 253–355 and 763–769. (Also forthcoming as a separate monograph, pub. Basil Blackwell.) "Identity and Necessity" in: *Identity and Individuation*, M. Munitz (ed.), New York University Press, 1971, pp. 135–164. Acquaintance with these papers is not a prerequisite for understanding the central puzzle of the present paper, but is helpful for understanding the theoretical background.

2 Frege gives essentially this example as the second footnote of "On Sense and Reference." For the "Who is . . . ?" to be applicable one must be careful to elicit from one's informant properties that he regards as defining the name and determining the referent, not mere well-known facts about the referent. (Of course this distinction may well seem fictitious, but it is central to the original Frege-Russell theory.)

3 For convenience Russell's terminology is assimilated to Frege's. Actually, regarding genuine or 'logically proper' names, Russell is a strict Millian: 'logically proper names' *simply* refer (to immediate objects of acquaintance). But, according to Russell, what are ordinarily called 'names' are not genuine, logically proper names, but disguised definite descriptions. Since Russell also regards definite descriptions as in turn disguised notation, he does not associate any 'senses' with descriptions, since they are not genuine singular terms. When all disguised notation is eliminated, the only singular terms remaining are logically proper names, for which no notion of 'sense' is required. When we speak of Russell as assigning 'senses' to names, we mean ordinary names and for convenience we ignore his view that the descriptions abbreviating them ultimately disappear on analysis.

On the other hand, the explicit doctrine that names are abbreviated definite descriptions is due to Russell. Michael Dummett, in his recent *Frege* (Duckworth and Harper and Row, 1973, pp. 110–111) denies that Frege held a description theory of senses. Although as far as I know Frege indeed makes no explicit statement to that effect, his examples of names conform to the doctrine, as Dummett acknowledges. Especially his 'Aristotle' example is revealing. He defines 'Aristotle' just as Russell would; it seems clear that in the case of a famous historical figure, the 'name' is indeed to be given by answering, in a uniquely specifying way, the 'who is' question. Dummett himself characterizes a sense as a "criterion . . . such that the referent of the name, if any, is whatever object satisfies that criterion." Since presumably the satisfaction of the criterion must be unique (so a unique referent is determined), doesn't this amount to defining names by unique satisfaction of properties, *i.e.*, by descriptions? *Perhaps* the point is that the property in question need not be expressible by a usual predicate of English, as might be plausible if the referent is one of the speaker's acquaintances rather than a historical figure. But I doubt that even Russell, father of the explicitly formulated description theory, ever meant to require that the description must always be expressible in (unsupplemented) English.

In any event, the philosophical community has generally understood Fregean senses in terms of descriptions, and we deal with it under this usual understanding. For present purposes this is more important than detailed historical issues. Dummett acknowledges (p. 111) that few substantive points are affected by his (allegedly) broader interpretation of Frege; and it would not seem to be relevant to the problems of the present paper.

4 See Frege's footnote in "On Sense and Reference" mentioned in note 2 above and

especially his discussion of 'Dr. Gustav Lauben' in "*Der Gedanke.*" (In the recent Geach-Stoothoff translation, "Thoughts," *Logical Investigations*, Oxford, Blackwell, 1977, pp. 11–12).

5 Russell, as a Millian with respect to genuine names, accepts this argument with respect to 'logically proper names.' For example — taking for the moment 'Cicero' and 'Tully' as 'logically proper names,' Russell would hold that if I judge that Cicero admired Tully, I am related to Cicero, Tully, and the admiration relation in a certain way: Since Cicero *is* Tully, I am related in exactly the same way to Tully, Cicero, and admiration; therefore I judge that Tully admired Cicero. Again, if Cicero *did* admire Tully, then according to Russell a single fact corresponds to all of 'Cicero admired Tully,' 'Cicero admired Cicero,' etc. Its constituent (in addition to admiration) is the man Cicero, taken, so to speak, twice.

Russell thought that 'Cicero admired Tully' and 'Tully admired Cicero' are in fact obviously not interchangeable. For him, this was one argument that 'Cicero' and 'Tully' are *not* genuine names, and that the Roman orator is no constituent of propositions (or 'facts,' or 'judgments') corresponding to sentences containing the name.

6 Given the arguments of Church and others, I do not believe that the formal mode of speech is synonymous with other formulations. But it can be used as a rough way to convey the idea of scope.

7 It may well be argued that the Millian view implies that proper names are *scopeless* and that for them the *de dicto–de re* distinction vanishes. This view has considerable plausibility (my own views on rigidity will imply something like this for *modal* contexts), but it need not be argued here either way: *de re* uses are simply not treated in the present paper.

Christopher Peacocke ("Proper Names, Reference, and Rigid Designation," in: *Meaning, Reference, and Necessity*, S. Blackburn (ed.), Cambridge, 1975; see Section I), uses what amounts to the equivalence of the *de dicto–de re* constructions in *all* contexts (or, put alternatively, the lack of such a distinction) to characterize the notion of rigid designation. I agree that for *modal* contexts, this is (roughly) equivalent to my own notion, also that for proper names Peacocke's equivalence holds for temporal contexts. (This is roughly equivalent to the 'temporal rigidity' of names.) I also agree that it is very plausible to extend the principle to all contexts. But, as Peacocke recognizes, this appears to imply a substitutivity principle for codesignative proper names in belief contexts, which is widely assumed to be false. Peacocke proposes to use Davidson's theory of intensional contexts to block this conclusion (the material in the 'that' clause is a separate sentence). I myself cannot accept Davidson's theory; but even if it were true, Peacocke in effect acknowledges that it does not really dispose of the difficulty (p. 127, first paragraph). (Incidentally, if Davidson's theory does block any inference to the transparency of belief contexts with respect to names, why does Peacocke assume without argument that it does not do so for modal contexts, which have a similar grammatical structure?) The problems are thus those of the present paper; until they are resolved I prefer at present to keep to my earlier more cautious formulation.

Incidentally, Peacocke hints a recognition that the received platitude — that codesignative names are not interchangeable in belief contexts — may not be so clear as is generally supposed.

8 The example comes from Quine, *Word and Object*, M.I.T. Press, 1960, p. 145. Quine's

conclusion that 'believes that' construed *de dicto* is opaque has widely been taken for granted. In the formulation in the text I have used the colon to emphasize that I am speaking of belief *de dicto*. Since, as I have said, belief *de dicto* will be our *only* concern in this paper, in the future the colon will usually be suppressed, and all 'believes that' contexts should be read *de dicto* unless the contrary is indicated explicitly.

9 In many writings Peter Geach has advocated a view that is nonMillian (he would say 'nonLockean') in that to each name a sortal predicate is attached by definition ('Geach,' for example, by *definition* names a man). On the other hand, the theory is not completely Fregean either, since Geach denies that any definite description that would identify the referent of the name among things of the same sort is analytically tied to the name. (See, for example, his *Reference and Generality*, Cornell, 1962, pp. 43–45.) As far as the present issues are concerned, Geach's view can fairly be assimilated to *Mill*'s rather than Frege's. For such ordinary names as 'Cicero' and 'Tully' will have both the same reference and the same (Geachian) sense (namely, that they are names of a man). It would thus seem that they ought to be interchangeable everywhere. (In *Reference and Generality*, Geach appears not to accept this conclusion, but the *prima facie* argument for the conclusion will be the same as on a purely Millian view.)

10 In an unpublished paper, Diana Ackerman urges the problem of substitutivity failures against the Millian view and, hence, against my own views. I believe that others may have done so as well. (I have the impression that the paper has undergone considerable revision, and I have not seen recent versions.) I agree that this problem is a considerable difficulty for the Millian view, and for the Millian *spirit* of my own views in "Naming and Necessity." (See the discussion of this in the text of the present paper.) On the other hand I would emphasize that there need be no *contradiction* in maintaining that names are *modally* rigid, and satisfy a substitutivity principle for modal contexts, while denying the substitutivity principle for belief contexts. The entire apparatus elaborated in "Naming and Necessity" of the distinction between epistemic and metaphysical necessity, and of giving a meaning and fixing a reference, was meant to show, among other things, that a Millian substitutivity doctrine for modal contexts can be maintained even if such a doctrine for epistemic contexts is rejected. "Naming and Necessity" never asserted a substitutivity principle for epistemic contexts.

It is even consistent to suppose that differing modes of (rigidly) fixing the reference is responsible for the substitutivity failures, thus adopting a position intermediate between Frege and Mill, on the lines indicated in the text of the present paper. "Naming and Necessity" may even perhaps be taken as suggesting, for some contexts where a conventional description rigidly fixes the reference ('Hesperus–Phosphorus'), that the mode of reference fixing is relevant to epistemic questions. I knew when I wrote "Naming and Necessity" that substitutivity issues in epistemic contexts were really very delicate, due to the problems of the present paper, but I thought it best not to muddy the waters further. (See notes 43–44.)

After this paper was completed, I saw Alvin Plantinga's paper "The Boethian Compromise," *The American Philosophical Quarterly* 15 (April, 1978): 129–138. Plantinga adopts a view intermediate between Mill and Frege, and cites substitutivity failures as a principal argument for his position. He also refers to a forthcoming paper by Ackerman. I have not seen this paper, but it probably is a descendant of the paper referred to above.

11 Here I use 'connotation' so as to imply that the associated properties have an *a priori* tie to the name, at least as rigid reference fixers, and therefore must be true of the referent (if it exists). There is another sense of 'connotation,' as in 'The Holy Roman Empire,' where the connotation need not be assumed or even believed to be true of the referent. In some sense akin to this, classicists and others with some classical learning may attach certain distinct 'connotations' to 'Cicero' and 'Tully.' Similarly, 'The Netherlands' may suggest low altitude to a thoughtful ear. Such 'connotations' can hardly be thought of as community-wide; many use the names unaware of such suggestions. Even a speaker aware of the suggestion of the name may not regard the suggested properties as true of the object; *cf.* 'The Holy Roman Empire.' A 'connotation' of this type neither gives a meaning nor fixes a reference.

12 Some might attempt to find a difference in 'sense' between 'Cicero' and 'Tully' on the grounds that "Cicero is called 'Cicero' " is trivial, but "Tully is called 'Cicero' " may not be. Kneale, and in one place (probably at least implicitly) Church, have argued in this vein. (For Kneale, see "Naming and Necessity," p. 283.) So, it may be argued, being called 'Cicero,' is part of the sense of the name 'Cicero,' but not part of that of 'Tully.'

I have discussed some issues related to this in "Naming and Necessity," pp. 283–286. (See also the discussions of circularity conditions elsewhere in "Naming and Necessity.") Much more could be said about and against this kind of argument; perhaps I will sometime do so elsewhere. Let me mention very briefly the following parallel situation (which may be best understood by reference to the discussion in "Naming and Necessity"). Anyone who understands the meaning of 'is called' and of quotation in English (and that 'alienists' is meaningful and grammatically appropriate), knows that "alienists are called 'alienists' " expresses a truth in English, even if he has no idea what 'alienists' means. He need *not* know that "psychiatrists are called 'alienists' " expresses a truth. None of this goes to show that 'alienists' and 'psychiatrists' are not synonymous, or that 'alienists' has *being called 'alienists'* as part of its meaning when 'psychiatrists' does not. Similarly for 'Cicero' and 'Tully.' There is no more reason to suppose that being so-called is part of the meaning of a name than of any other word.

13 A view follows Frege and Russell on this issue even if it allows each speaker to associate a cluster of descriptions with each name, provided that it holds that the cluster varies from speaker to speaker and that variations in the cluster are variations in idiolect. Searle's view thus is Frege-Russellian when he writes in the concluding paragraph of "Proper Names" (*Mind* 67 (1958): 166–173), " 'Tully = Cicero' would, I suggest, be analytic for most people; the same descriptive presuppositions are associated with each name. But of course if the descriptive presuppositions were different it might be used to make a synthetic statement."

14 Though here I use the jargon of propositions, the point is fairly insensitive to differences in theoretical standpoints. For example, on Davidson's analysis, I would be asserting (roughly) that many are unaware-of-the-content-of the following *utterance* of mine: Cicero is Tully. This would be subject to the same problem.

15 Benson Mates, "Synonymity," *University of California Publications in Philosophy* 25 (1950): 201–226; reprinted in: *Semantics and the Philosophy of Language*, L. Linsky (ed.), University of Illinois Press, 1952. (There was a good deal of subsequent discussion. In Mates's original paper the point is made almost parenthetically.) Actually, I think that

Mates's problem has relatively little force against the argument we are considering for the Fregean position. Mates's puzzle in no way militates against some such principle as: If one word is synonymous with another, then a sufficiently reflective speaker subject to no linguistic inadequacies or conceptual confusions who sincerely assents to a simple sentence containing the one will also (sincerely) assent to the corresponding sentence with the other in its place.

It is surely a crucial part of the present 'Fregean' argument that codesignative names may have distinct 'senses,' that a speaker may assent to a simple sentence containing one and deny the corresponding sentence containing the other, even though he is *guilty of no conceptual or linguistic confusion, and of no lapse in logical consistency*. In the case of two straightforward synonyms, this is not so.

I myself think that Mates's argument is of considerable interest, but that the issues are confusing and delicate and that, if the argument works, it probably leads to a paradox or puzzle rather than to a definite conclusion. (See also notes 23, 28, and 46.)

16 "Naming and Necessity," pp. 291 (bottom)–293.

17 Recall also note 12.

18 Some philosophers stress that names are not *words* of a language, or that names are not *translated* from one language to another. (The phrase 'common currency of our common language' was meant to be neutral with respect to any such alleged issue.) Someone may use 'Mao Tse-Tung,' for example, in English, though he knows not one word of Chinese. It seems hard to deny, however, that "*Deutschland*," "*Allemagne*," and "Germany," are the German, French, and English names of a single country, and that one translates a French sentence using "*Londres*" by an English sentence using "London." Learning these facts *is* part of learning German, French, and English.

It would appear that *some* names, especially names of countries, other famous localities, and some famous people *are* thought of as part of a language (whether they are called 'words' or not is of little importance). Many other names are not thought of as part of a language, especially if the referent is not famous (so the notation used is confined to a limited circle), or if the same name is used by speakers of all languages. As far as I can see, it makes little or no *semantic* difference whether a particular name is thought of as part of a language or not. Mathematical notation such as '$<$' is also ordinarily not thought of as part of English, or any other language, though it is used in combination with English words in sentences of mathematical treatises written in English. (A French mathematician can use the notation though he knows not one word of English.) 'Is less than,' on the other hand, *is* English. Does this difference have any semantic significance?

I will speak in most of the text as if the names I deal with are part of English, French, etc. But it matters little for what I say whether they are thought of as parts of the language or as adjuncts to it. And one need not say that a name such as '*Londres*' is 'translated' (if such a terminology suggested that names have 'senses,' I too would find it objectionable), as long as one acknowledges that *sentences* containing it are properly translated into English using 'London.'

19 By saying that names are transparent in a context, I mean that codesignative names are interchangeable there. This is a deviation for brevity from the usual terminology, according to which the *context* is transparent. (I use the usual terminology in the paper also.)

20 But we must use the term 'sense' here in the sense of 'that which fixes the reference,' not 'that which gives the meaning,' otherwise we shall run afoul of the rigidity of proper names. If the source of a chain for a certain name is in fact a given object, we use the name to designate that object even when speaking of counterfactual situations in which some *other* object originated the chain.

21 The point is that, according to the doctrine of "Naming and Necessity," when proper names are transmitted from link to link, even though the beliefs about the referent associated with the name change radically, the change is not to be considered a linguistic change, in the way it *was* a linguistic change when 'villain' changed its meaning from 'rustic' to 'wicked man.' As long as the reference of a name remains the same, the associated beliefs about the object may undergo a large number of changes without these changes constituting a change in the language.

If Geach is right, an appropriate sortal must be passed on also. But see footnote 58 of "Naming and Necessity."

22 Similar appropriate restrictions are assumed below for the strengthened disquotational principle and for the principle of translation. Ambiguities need not be excluded if it is tacitly assumed that the sentence is to be understood in one way in all its occurrences. (For the principle of translation it is similarly assumed that the translator matches the *intended* interpretation of the sentence.) I do not work out the restrictions on indexicals in detail, since the intent is clear.

Clearly, the disquotational principle applies only to *de dicto*, not *de re*, attributions of belief. If someone sincerely assents to the near triviality "The tallest foreign spy is a spy," it follows that he believes that: the tallest foreign spy is a spy. It is well known that it does *not* follow that he believes, *of* the tallest foreign spy, that he is a spy. In the latter case, but not in the former, it would be his patriotic duty to make contact with the authorities.

23 What if a speaker assents to a sentence, but fails to assent to a synonymous assertion? Say, he assents to "Jones is a doctor," but not to "Jones is a physician." Such a speaker either does not understand one of the sentences normally, or he should be able to correct himself "on reflection." As long as he confusedly assents to 'Jones is a doctor' but not to 'Jones is a physician,' we *cannot* straightforwardly apply disquotational principles to conclude that he does or does not believe that Jones is a doctor, because his assent is not "reflective."

Similarly, if someone asserts, "Jones is a doctor but not a physician," he should be able to recognize his inconsistency without further information. We have formulated the disquotational principles so they need not lead us to attribute belief as long as we have grounds to suspect conceptual or linguistic confusion, as in the cases just mentioned.

Note that if someone says, "Cicero was bald but Tully was not," there need be *no* grounds to suppose that he is under *any* linguistic or conceptual confusion.

24 This should not be confused with the question whether the speaker simultaneously believes *of* a given object, both that it has a certain property and that it does not have it. Our discussion concerns *de dicto* (notional) belief, not *de re* belief.

I have been shown a passage in Aristotle that appears to suggest that *no one* can really believe both of two explicit contradictories. If we wish to use the *simple* disquotational principle as a test for disbelief, it suffices that this be true of *some* individuals, after reflection, who are simultaneously aware of both beliefs, and have sufficient logical acumen and

respect for logic. Such individuals, if they have contradictory beliefs, will be shaken in one or both beliefs after they note the contradiction. For such individuals, sincere reflective assent to the negation of a sentence implies disbelief in the proposition it expresses, so the test in the text applies.

25 For example, in translating a historical report into another language, such as, "Patrick Henry said, 'Give me liberty or give me death!' " the translator may well translate the quoted material attributed to Henry. He translates a presumed truth into a falsehood, since Henry spoke English; but probably his reader is aware of this and is more interested in the content of Henry's utterance than in its exact words. Especially in translating fiction, where truth is irrelevant, this procedure is appropriate. But some objectors to Church's 'translation argument' have allowed themselves to be misled by the practice.

26 To state the argument precisely, we need in addition a form of the Tarskian disquotation principle for truth: For each (French or English) replacement for '*p*,' infer " '*p*' is true" from "*p*," and conversely. (Note that " 'p' is true" becomes an English sentence even if '*p*' is replaced by a French sentence.) In the text we leave the application of the Tarskian disquotational principle tacit.

27 I gather that Burali-Forti originally thought he had 'proved' that the ordinals are not linearly ordered, reasoning in a manner similar to our topologist. Someone who heard the present paper delivered told me that König made a similar error.

28 It is not possible, in this case, as it is in the case of the man who assents to "Jones is a doctor" but not to "Jones is a physician," to refuse to apply the disquotational principle on the grounds that the subject must lack proper command of the language or be subject to some linguistic or conceptual confusion. As long as Pierre is unaware that 'London' and '*Londres*' are codesignative, he need not lack appropriate linguistic knowledge, nor need he be subject to any linguistic or conceptual confusion, when he affirms '*Londres est jolie*' but denies 'London is pretty.'

29 The 'elimination' would be most plausible if we believed, according to a Russellian epistemology, that all my language, when written in unabbreviated notation, refers to constituents with which I am 'acquainted' in Russell's sense. Then no one speaks a language intelligible to anyone else; indeed, no one speaks the same language twice. Few today will accept this.

A basic consideration should be stressed here. Moderate Fregeans attempt to combine a roughly Fregean view with the view that names are part of our common language, and that our conventional practices of interlinguistic translation and interpretation are correct. The problems of the present paper indicate that it is very difficult to obtain a requisite socialized notion of sense that will enable such a program to succeed. Extreme Fregeans (such as Frege and Russell) believe that in general names are peculiar to idiolects. They therefore would accept no general rule translating '*Londres*' as 'London,' nor even translating one person's use of 'London' into another's. However, if they follow Frege in regarding senses as 'objective,' they must believe that in principle it makes sense to speak of two people using two names in their respective idiolects with the same sense, and that there must be (necessary and) sufficient conditions for this to be the case. If these conditions for sameness of sense are satisfied, translation of one name into the other is legitimate, otherwise not. The present considerations (and the extension of these below to natural kind and related terms), however, indicate that the notion of sameness of sense, if it is to be explicated in

terms of sameness of identifying properties and if these properties are themselves expressed in the languages of the two respective idiolects, presents interpretation problems of the same type presented by the names themselves. Unless the Fregean can give a method for identifying sameness of sense that is free of such problems, he *has no sufficient conditions for sameness of sense, nor for translation to be legitimate*. He would therefore be forced to maintain, contrary to Frege's intent, that not only in practice do few people use proper names with the same sense but that *it is in principle meaningless to compare senses*. A view that the identifying properties used to define senses should always be expressible in a Russellian language of 'logically proper names' would be one solution to this difficulty but involves a doubtful philosophy of language and epistemology.

30 If any reader finds the term 'translation' objectionable with respect to names, let him be reminded that all I mean is that French sentences containing '*Londres*' are uniformly translated into English with 'London.'

31 The paradox would be blocked if we required that they define the names by the same properties expressed in the same words. There is nothing in the motivation of the classical description theories that would justify this extra clause. In the present case of French and English, such a restriction would amount to a decree that neither '*Londres*,' nor any other conceivable French name, could be translated as 'London.' I deal with this view immediately below.

32 Word salads of two languages (like ungrammatical 'semisentences' of a single language) need not be unintelligible, though they are makeshifts with no fixed syntax. "If God did not exist, Voltaire said, *il faudrait l'inventer*." The meaning is clear.

33 Had we said, "Pierre believes that the country he calls '*Angleterre*' is a monarchy," the sentence would be English, since the French word would be mentioned but not used. But for this very reason we would not have captured the sense of the French original.

34 Under the influence of Quine's *Word and Object*, some may argue that such conclusions are not inevitable: perhaps he will translate '*médecin*' as 'doctor stage,' or 'undetached part of a doctor'! If a Quinean skeptic makes an empirical prediction that such reactions from bilinguals as a matter of fact can occur, I doubt that he will be proved correct. (I don't know what Quine would think. But see *Word and Object*, p. 74, first paragraph.) On the other hand, if the translation of 'médecin' as 'doctor' rather than 'doctor part' in this situation *is*, empirically speaking, inevitable, then even the advocate of Quine's thesis will have to admit that there is something special about one particular translation. The issue is not crucial to our present concerns, so I leave it with these sketchy remarks. But see also note 36.

35 Putnam gives the example of elms and beeches in "The Meaning of 'Meaning' " (in: *Language, Mind, and Knowledge*, Minnesota Studies in the Philosophy of Science 7; also reprinted in Putnam's *Collected Papers*). See also Putnam's discussion of other examples on pp. 139–143; also my own remarks on 'fool's gold,' tigers, etc., in "Naming and Necessity," pp. 316–323.

36 It is unclear to me how far this can go. Suppose Pierre hears English spoken only in England, French in France, and learns both by direct method. (Suppose also that no one else in each country speaks the language of the other.) Must he be sure that 'hot' and '*chaud*' are coextensive? In practice he certainly would. But suppose somehow his experience is consistent with the following bizarre — and of course, false! — hypothesis: England and France differ atmospherically so that human bodies are affected very differently by their interaction with the surrounding atmosphere. (This would be more plausible if France were on another planet.) In particular, within reasonable limits, things that feel cold in one of the countries feel hot in the other, and *vice versa*. Things don't change their *temperature* when moved from England to France, they just *feel* different because of their effects on human physiology. Then '*chaud*,' in French, would be true of the things that are called 'cold' in English! (Of course the present discussion is, for space, terribly compressed. See also the discussion of 'heat' in "Naming and Necessity." We are simply creating, for the physical property 'heat,' a situation analogous to the situation for natural kinds in the text.)

If Pierre's experiences were arranged somehow so as to be consistent with the bizarre hypothesis, and he somehow came to believe it, he might simultaneously assent to '*C'est chaud*' and 'This is cold' without contradiction, even though he speaks French and English normally in each country separately.

This case needs much more development to see if it can be set up in detail, but I cannot consider it further here. Was I right in assuming in the text that the difficulty could not arise for '*médecin*' and 'doctor'?

37 One might argue that Peter and we do speak different dialects, since in Peter's idiolect 'Paderewski' is used ambiguously as a name for a musician and a statesman (even though these are in fact the same), while in our language it is used unambiguously for a musician-statesman. The problem then would be whether Peter's dialect can be translated homophonically into our own. Before he hears of 'Paderewski-the-statesman,' it would appear that the answer is affirmative for his (then unambiguous) use of 'Paderewski,' since he did not differ from anyone who happens to have heard of Paderewski's musical achievements but not of his statesmanship. Similarly for his later use of 'Paderewski,' if we ignore his earlier use. The problem is like Pierre's, and is essentially the same whether we describe it in terms of whether Peter satisfies the condition for the disquotational principle to be applicable, or whether homophonic translation of his dialect into our own is legitimate.

38 D. Davidson, "On Saying That," in: *Words and Objections*, D. Davidson and J. Hintikka (eds.), Dordrecht, Reidel, 1969, p. 166.

39 In *Word and Object*, p. 221, Quine advocates a second level of canonical notation, "to dissolve verbal perplexities or facilitate logical deductions," admitting the propositional attitudes, even though he thinks them "baseless" idioms that should be excluded from a notation "limning the true and ultimate structure of reality."

40 In one respect the considerations mentioned above on natural kinds show that Quine's translation apparatus is insufficiently skeptical. Quine is sure that the native's *sentence* "Gavagai!" should be translated "Lo, a rabbit!", provided that its affirmative and negative stimulus meanings for the native match those of the English sentence for the Englishman; skepticism sets in only when the linguist proposes to translate the *general term* 'gavagai' as 'rabbit' rather than 'rabbit stage,' 'rabbit part,' and the like. But there is another possibility that is independent of (and less bizarre than) such skeptical alternatives. In the geographical area inhabited by the natives, there may be a species indistinguishable to the nonzoologist from rabbits but forming a distinct species. Then the 'stimulus meanings,' in Quine's sense, of 'Lo, a rabbit!' and 'Gavagai!' may well be identical (to nonzoologists), especially if the ocular irradiations in question do not include a specification of the geographical locality.

('Gavagais' produce the same ocular irradiation patterns as rabbits.) Yet 'Gavagai!' and 'Lo, a rabbit!' are hardly synonymous; on typical occasions they will have opposite truth values.

I believe that the considerations about names, let alone natural kinds, emphasized in "Naming and Necessity" go against any simple attempt to base interpretation solely on maximizing agreement with the affirmations attributed to the native, matching of stimulus meanings, etc. The 'Principle of Charity' on which such methodologies are based was first enunciated by Neil Wilson in the special case of proper names as a formulation of the cluster-of-descriptions theory. The argument of "Naming and Necessity" is thus directed against the simple 'Principle of Charity' for that case.

41 Geach introduced the term 'Shakespearean' after the line, "a rose / By any other name, would smell as sweet."

Quine seems to define 'referentially transparent' contexts so as to imply that coreferential names and definite descriptions must be interchangeable *salva veritate*. Geach stresses that a context may be 'Shakespearean' but not 'referentially transparent' in this sense.

42 Generally such cases may be slightly less watertight than the 'London'–'*Londres*' case. '*Londres*' just is the French version of 'London,' while one cannot quite say that the same relation holds between '*Ashkenaz*' and '*Germaniah*.' Nevertheless:

(a) Our standard practice in such cases is to translate both names of the first language into the single name of the second.

(b) Often no nuances of 'meaning' are discernible differentiating such names as '*Ashkenaz*' and '*Germaniah*,' such that we would not say either that Hebrew would have been impoverished had it lacked one of them (or that English is improverished because it has only one name for Germany), any more than a language is impoverished if it has only one word corresponding to 'doctor' and 'physician.' Given this, it seems hard to condemn our practice of translating both names as 'Germany' as 'loose'; in fact, it would seem that Hebrew just has two names for the same country where English gets by with one.

(c) Any inclinations to avoid problems by declaring, say, the translation of '*Ashkenaz*' as 'Germany' to be loose should be considerably tempered by the discussion of analogous problems in the text.

43 In spite of this official view, perhaps I will be more assertive elsewhere.

In the case of 'Hesperus' and 'Phosphorus' (in contrast to 'Cicero' and 'Tully'), where there is a case for the existence of conventional community-wide 'senses' differentiating the two — at least, two distinct modes of 'fixing the reference of two rigid designators' — it is more plausible to suppose that the two names are definitely not interchangeable in belief contexts. According to such a supposition, a belief that Hesperus is a planet is a belief that a certain heavenly body, rigidly picked out as seen in the evening in the appropriate season, is a planet; and similarly for Phosphorus. One may argue that translation problems like Pierre's will be blocked in this case, that '*Vesper*' must be translated as 'Hesperus,' not as 'Phosphorus.' As against this, however, two things:

(a) We should remember that sameness of properties used to fix the reference does *not* appear to guarantee in general that paradoxes will not arise. So one may be reluctant to adopt a solution in terms of reference-fixing properties for this case if it does not get to the heart of the general problem.

(b) The main issue seems to me here to be — how essential is a particular mode of fixing the reference to a correct learning of the name? If a parent, aware of the familiar identity, takes a child into the fields in the morning and says (pointing to the morning star) "That is called 'Hesperus,' " has the parent mistaught the language? (A parent who says, "Creatures with kidneys are called 'cordates,' definitely has mistaught the language, even though the statement is extensionally correct.) To the extent that it is *not* crucial for correct language learning that a particular mode of fixing the reference be used, to that extent there is no 'mode of presentation' differentiating the 'content' of a belief about 'Hesperus' from one about 'Phosphorus.' I am doubtful that the original method of fixing the reference *must* be preserved in transmission of the name.

If the mode of reference fixing *is* crucial, it can be maintained that otherwise identical beliefs expressed with 'Hesperus' and with 'Phosphorus' have definite differences of 'content,' at least in an epistemic sense. The conventional ruling against substitutivity could thus be maintained without qualms for some cases, though not as obviously for others, such as 'Cicero' and 'Tully.' But it is unclear to me whether even 'Hesperus' and 'Phosphorus' do have such conventional 'modes of presentation.' I need not take a definite stand, and the verdict may be different for different particular pairs of names. For a brief related discussion, see "Naming and Necessity," p. 331, first paragraph.

44 However, some earlier formulations expressed disquotationally such as "It was once unknown that Hesperus is Phosphorus" are questionable in the light of the present paper (but see the previous note for this case). I was aware of this question by the time "Naming and Necessity" was written, but I did not wish to muddy the waters further than necessary at that time. I regarded the distinction between epistemic and metaphysical necessity as valid in any case and adequate for the distinctions I wished to make. The considerations in this paper are relevant to the earlier discussion of the 'contingent *a priori*' as well; perhaps I will discuss this elsewhere.

45 According to Russell, definite descriptions are not genuine singular terms. He thus would have regarded any concept of 'referential opacity' that includes definite descriptions as profoundly misleading. He also maintained a substitutivity principle for 'logically proper names' in belief and other attitudinal contexts, so that for him belief contexts were as 'transparent,' in any philosophically decent sense, as truth-functional contexts.

Independently of Russell's views, there is much to be said for the opinion that the question whether a context is 'Shakespearean' is more important philosophically — even for many purposes for which Quine invokes his own concept — than whether it is 'referentially opaque.'

46 I will make some brief remarks about the relation of Benson Mates's problem (see note 15) to the present one. Mates argued that such a sentence as (*)'Some doubt that all who believe that doctors are happy believe that physicians are happy,' may be true, even though 'doctors' and 'physicians' are synonymous, and even though it would have been false had 'physicians' been replaced in it by a second occurrence of 'doctors.' Church countered that (*) could not be true, since its translation into a language with only one word for doctors (which would translate both 'doctors' and 'physicians') would be false. If *both* Mates's and Church's intuitions were correct, we might get a paradox analogous to Pierre's.

Applying the principles of translation and disquotation to Mates's puzzle, however,

involves many more complications than our present problem. First, if someone assents to 'Doctors are happy,' but refuses assent to 'Physicians are happy,' *prima facie* disquotation does not apply to him since he is under a linguistic or conceptual confusion. (See note 23.) So there are as yet no grounds, merely because this happened, to doubt that all who believe that doctors are happy believe that physicians are happy.

Now suppose someone assents to 'Not all who believe that doctors are happy believe that physicians are happy.' What is the source of his assent? If it is failure to realize that 'doctors' and 'physicians' are synonymous (this was the situation Mates originally envisaged), then he is under a linguistic or conceptual confusion, so disquotation does not clearly apply. Hence we have no reason to conclude from this case that (*) is true. Alternatively, he may realize that 'doctors' and 'physicians' are synonymous; but he applies disquotation to a man who assents to 'Doctors are happy' but not to 'Physicians are happy,' ignoring the caution of the previous paragraph. Here he is not under a simple linguistic confusion (such as failure to realize that 'doctors' and 'physicians' are synonymous), but he appears to be under a deep conceptual confusion (misapplication of the disquotational principle). Perhaps, it may be argued, he misunderstands the 'logic of belief.' Does his conceptual confusion mean that we cannot straightforwardly apply disquotation to his utterance, and that therefore we cannot conclude from his behavior that (*) is true? I think that, although the issues are delicate, and I am not at present completely sure what answers to give, there is a case for an affirmative answer. (Compare the more extreme case of someone who is so confused that he thinks that someone's *dissent* from 'Doctors are happy' implies that he believes that doctors are happy. If someone's utterance, 'Many believe that doctors are happy,' is based on such a misapplication of disquotation, surely we in turn should not apply disquotation to it. The utterer, at least in this context, does not really know what 'belief' means.)

I do *not* believe the discussion above ends the matter. Perhaps I can discuss Mates's problem at greater length elsewhere. Mates's problem is perplexing, and its relation to the present puzzle is interesting. But it should be clear from the preceding that Mates's argument involves issues even more delicate than those that arise with respect to Pierre. First, Mates's problem involves delicate issues regarding iteration of belief contexts, whereas the puzzle about Pierre involves the application of disquotation only to affirmations of (or assents to) *simple* sentences. More important, Mates's problem would not arise in a world where no one ever was under a linguistic or a conceptual confusion, no one ever thought anyone else was under such a confusion, no one ever thought anyone ever thought anyone was under such a confusion, and so on. It is important, both for the puzzle about Pierre and for the Fregean argument that 'Cicero' and 'Tully' differ in 'sense,' that they would still arise in such a world. They are entirely free of the delicate problem of applying disquotation to utterances directly or indirectly based on the existence of linguistic confusion. See notes 15 and 28, and the discussion in the text of Pierre's logical consistency.

Another problem discussed in the literature to which the present considerations may be relevant is that of 'self-consciousness,' or the peculiarity of 'I.' Discussions of this problem have emphasized that 'I,' even when Mary Smith uses it, is not interchangeable with 'Mary Smith,' nor with any other conventional singular term designating Mary Smith. If she is 'not aware that she is Mary Smith,' she may assent to a sentence with 'I,' but dissent from the corresponding sentence with 'Mary Smith.' It is quite possible that any attempt

to clear up the logic of all this will involve itself in the problem of the present paper. (For this purpose, the present discussion might be extended to demonstratives and indexicals.)

The writing of this paper had partial support from a grant from the National Science Foundation, a John Simon Guggenheim Foundation Fellowship, a Visiting Fellowship at All Souls College, Oxford, and a sabbatical leave from Princeton University. Various people at the Jerusalem Encounter and elsewhere, who will not be enumerated, influenced the paper through discussion.