Dissonant Notes on the Theory of Reference[1]

STEPHEN P. STICH
UNIVERSITY OF MICHIGAN

1. Professor Quine would divide the problems of semantics into two distinct provinces: *the theory of meaning* and *the theory of reference*. Central to the former are the notions of *meaning, synonymy, significance, analyticity* and *entailment*. Central to the latter, the concepts of *naming, truth, denotation (or truth of), extension* and *values of a variable* ([3], p. 130). The principle behind the division is not obvious, but the enumeration of the basic concepts of each theory renders the division itself workably clear.

The two theories are not equal in Quine's esteem. "The theory of meaning," on his view, "is in a worse state than the theory of reference." Its notions, by comparison to those of the theory of reference, are "foggy and mysterious."

I will contend that Quine's optimism about the theory of reference is incompatible with his pessimism about the theory of meaning. For, on Quine's own account, the problems that discourage him about the theory of meaning beset the theory of reference as well. And of the three arguments Quine advances to show the theory of reference better off than the theory of meaning, two are unsound and the third is in conflict with his further views on reference.

2. Quine views "language as the complex of present dispositions to verbal behavior" ([4], p. 27). This formulation invites quibbles over 'verbal behavior' and quarrels over 'dispositions', but the general view is one I share. If a theory of language is not, in some suitably vague sense, a theory of verbal behavior, it is hard to imagine what it is a theory of.

[1] I am indebted to Paul Benacerraf, Wilbur D. Hart, Jaegwon Kim and John R. Wallace for their helpful comments on earlier drafts of this paper.

Both the theory of meaning and the theory of reference are alleged to be part of a theory of language. Thus Quine would have the concepts of these theories introduced into the theory Carnap calls *pragmatics*—an empirical discipline dealing (roughly speaking) with human verbal behavior. Pragmatic concepts make explicit reference to the speaker. Thus, to start, what we want explained are 'S is an analytic statement for person P', 'expressions e and e′ are synonymous for person P', 'S is true for person P' and so on, for variable 'S', 'e', 'e′' and 'P'. Derivatively, we may substitute 'in language L' for 'for person P' and talk collectively of the speakers of a language. This requires some independent explanation of 'person P speaks language L'. (Much more need be said on all this, but not here.)

To explain these concepts we might show how they can be defined in terms of verbal behavior. Less restrictively, we can show the part they play in the pragmatic theory as a whole. Though the question of what to allow in an explanation "in terms of verbal behavior" is complicated and not uncontroversial, two general points will, I think, be readily accepted. First we must know for each pragmatic claim of the sorts specified above what sorts of (verbal) behavior would count as evidence for or against it. Second, if the theory put forward requires that exactly one of several such statements must be true, then we must have explained what evidence would count for each alternative and against all the others. If all verbal behavior counts equally for each, then the theory has drawn "a distinction without a difference." Further requirements will surely have to be placed on the introduction of concepts in pragmatics, but these two are clearly necessary if we are to understand the concepts at all.

Quine's qualms about the theory of meaning can be traced to the absence of an adequate behavioral explication for its central concepts. While theories of meaning commonly insist, for example, that any pair of unambiguous predicates must be either synonymous or heteronymous, they are uninformative on the sorts of evidence that, in crucial cases, would count for one hypothesis and against the other. Thus, what will distinguish a pair of synonymous predicates from a heteronymous pair which the speaker firmly believes to be true of the same things? Quine can find no satisfactory answer. He suggests the distinction itself is illusory ([4], Sec. 11).

What, now, of the theory of reference? Do its central notions prove behaviorally more perspicuous? On Quine's view they do not.

For, as he has argued at least four times ([4], Ch. II; [5]; [6]; [7]), there is no behavioral evidence that would count for 'In Jungle 'gavagai' is true of rabbits' and against the competing hypotheses with 'rabbits' replaced by 'undetached parts of rabbits' or 'temporal segments of rabbits'. Thus "reference itself proves behaviorally inscrutable" ([7], p. 191).

3. How then may we explain Quine's optimism about the theory of reference? It can, I think, be traced to Tarski's work on the concept of truth. Quine marshals three arguments, each suggested by Tarski's work and each aimed at showing the notions of the theory of reference "very much less foggy and mysterious than the notions belonging to the theory of meaning" ([3], pp. 137–8). The three together, Quine feels, endow the terms of the theory of reference "with a high enough degree of intelligibility so that we are not likely to be averse to using the idiom" ([3], p. 138). It is these arguments I propose to challenge.

4. Quine grants that Tarski's work offers "no . . . single definition of 'true-in-L' for variable 'L' " ([3], p. 138). Nor does it provide us with a definition of 'true-in-L' where L is some particular natural language. Indeed, for those languages which Tarski calls "universal" the task of constructing such a definition is known to be impossible, for it is in such languages that the "semantic paradoxes" arise. But while we do not have general definitions of the concepts of the theory of reference, Quine holds that Tarski's work provides us with a "clue" of considerable value. The clue is to be found in the "paradigms" inspired by Tarski's convention T.

(1) '———' is true-in-L if and only if ———.
(2) '———' is true-in-L of every ——— thing and nothing else.
(3) '———' names-in-L ——— and nothing else.

(Cf. [3], p. 135; for Tarski's convention T, cf. [8], pp. 187–8.) The paradigms are not definitions. But, and this is Quine's first argument, the paradigms at least tell us what it would be to get the desired definitions right.

Consider first the case in which the meta-language, ML, we use to talk about truth in an object language, L, *contains* L. A definition of 'true-in-L' must have as a consequence every statement of ML formed by placing any one statement of L in the blanks of (1). What is more, the paradigm leaves no ambiguity as to the extension

of the concept of truth in L. For suppose we have two different interpretations of 'true-in-L', say 'true$_1$-in-L' and 'true$_2$-in-L'. Let $(1)_1$ and $(1)_2$ be the result of substituting these for 'true-in-L' in (1). Then "from $[(1)_1]$ and $[(1)_2]$ it follows logically that

'———' is true$_1$-in-L if and only if '———' is true$_2$-in-L

no matter what statement of L we write for '———'. Thus truth$_1$-in-L and truth$_2$-in-L coincide. Similar reasoning works for [(2)] and [(3)]" ([3], p. 136).

In addition to providing a criterion of adequacy for proposed definitions, the paradigm, on Quine's view, offers yet another clue of considerable value, and with it a second argument in favor of the concepts of the theory of reference. (1) "serves to endow 'true-in-L' . . . with every bit as much clarity, in particular applications, as is enjoyed by the particular expressions of L to which we apply [it]. Attribution of truth to 'Snow is white', for example, is every bit as clear to us as attribution of whiteness to snow" ([3], p. 138).

So much for the case where ML includes L as a part. But now what of those object languages *not* included in our chosen meta-language? Suppose, for example, our meta-language is a variant of ordinary English, shorn of semantic terms and provided with some systematic way of naming expressions—call it *meta-English* (ME). What are we to make of 'true-in-German' or 'true-in-Black-Thai'? Here Quine offers an imaginative proposal. Putting a German statement for the blank in (1) would yield nonsense—a sentence part in English and part in German. But suppose we take as our meta-language not ME, our tidied English, but rather the result of pooling ME with our object language, German. In this new meta-language our recently imagined substitution in (1) makes perfectly good sense.

This feat of "pooling" the object language with ME to produce an adequate meta-language is not at all so clear as Quine's quick talk of a "composite language" suggests. To restrict our criticism, for the moment, to small points, what are the rules for forming noun phrases in the composite German-English? Can we use a German article and an English noun, and if so, what is the gender of the noun? Just how much is packed into this talk of pooling will emerge in due time. But rather than continue along this line, let us return to Quine's arguments granting, for the time being, that some general rules can be laid down for the bothersome details of pooling languages.

Observe now what this device of pooling the object language and ME to form a meta-language has accomplished. We still cannot see our way clear to a definition of 'true-in-L' in our ML. But, nonetheless there has been substantial progress. For we can now give the sort of criterion of adequacy for definitions of 'true-in-L' that was discussed four paragraphs back. Further, we can render every attribution of truth to a statement of the object language every bit as clear as the statement itself. And all of this has been accomplished without treading in the mire of the theory of meaning. Of course the statements of the object language, L, and thus those of our merged ML, are incomprehensible unless we *understand* L. But understanding a language is a notion we can accept without dabbling in meanings. If someone should question our talk of truth in L, we explain it to him as follows: Go out and learn L; use any method you find appropriate. When you understand the statements of L you will understand the result of substituting these statements in (1). Now what we seek in a definition of truth-in-L is a definition that will have each of these substitutions as a consequence. And attributing truth to a statement of L should be as clear to you as the statement itself.

The force of these arguments appears when we contrast truth to analyticity. For suppose we try the same device on the latter notion. We can come to understand L much as any immigrant comes to understand the language of his adopted land. With this understanding 'true-in-L' acquires the degree of intelligibility the paradigm provides. But even when we are as comfortable as a native in L, what do we understand of 'analytic-in-L'? "We have no clue comparable in value to [(1)]" ([3], p. 138), no idea of what it would be to get a definition of 'analytic-in-L' right, nor any idea of the import of individual attributions of analyticity to statements of L. Rather "definition of analytic-in-L for each L seems . . . to be a project unto itself. The most evident principle of unification, linking analyticity-in-L for one choice of L with analyticity-in-L for another choice of L, is the joint use of the syllables 'analytic'" ([3], p. 138).

Here, then, are two of the arguments Quine offers in defence of the theory of reference. I find them ingenious and, at first blush, compelling. I also think they are wrong.

5. Quine has been depicted as claiming two separate virtues for (1). First it provides a criterion of adequacy for any proposed definition, insuring that any two definitions of truth in a language will

at least pick out the same statements. Second, in the absence of a definition, it serves to clarify any particular attribution of truth to a statement. Let me tackle these one at a time starting with the second.

To prepare the groundwork for my attack, let us dredge up some of the confusions (1) has engendered. The paradigm, and its inspiration, Tarski's convention T, have been the source of much philosophical perplexity. This perplexity is evidenced in an article by Max Black ([1]). Black is troubled by the fact that in seeing Tarski's convention and following through his construction in a sample language, we "seem to *understand* Tarski's procedure. . . . We feel we *understand* the definition" ([1], p. 102). That is, we seem to grasp the "principle" of the definition. What is troubling in all this is that when we try to *say* what it is we understand, it seems to elude us. Thus, to use Quine's formulation of the paradigm, what precisely does (1) tell us? A familiar bit of nonsense arises from trying to treat the blanks of (1) as variables. One move along these lines would try to capture the "point" of (1) by:

(4) (x) 'x' is true-in-L if and only if x.

But this is a double muddle. On the left side it is attributing truth to the 24th letter of the alphabet, not to a statement. And on the right side, what are we to take the variables to be ranging over? If statements, then we must replace the variables with *names* of statements. We might do somewhat better by taking the variable to range over propositions, the intensional entities that statements are sometimes said to name. But this step back into the realm of intensions still leaves the left side of (4) attributing truth to a letter of the alphabet. Also, on currently fashionable accounts, statements name truth values, not propositions.

Clearly (4) is a failure as an attempt to capture the point of (1). But if not (4), what? The answer becomes clear when we realize that what is claimed to hold is the result of putting any one statement of the object language for the blanks of (1). This resulting statement is part of the meta-language (assuming the object language to be contained within the meta-language). So to say it is true, we must go one step higher, to the meta-meta-language (MML). Thus Quine writes:

> In general, if language L (for example, German) is contained in language L′ (for example, German-English), so that L′ is simply L or else L plus some supplementary vocabulary or grammatical

> constructions, and if the portions, at least, of English usage which figure in [(1)] above (apart from the blanks) are part of L' then the result of putting any one statement of L for the blanks in [(1)] *is true in L'*. ([3], p. 135—emphasis mine.)

A plausible rendition of this would be the following in MML:

(5) For all x and y, if x is a statement of L and y is the quote name of x in ML, then the result of substituting y for 'z*' and x for 'z' in 'z*' is true-in-L if and only if z' is true-in-ML.

(This is a variation on a suggestion by P. T. Geach in [2].)

But now there is something startling in both Quine's remark and our gloss, for *both of them use the notion of truth-in-the-meta-language.* Both statements belong to MML and each uses 'true-in-ML'. What the paradigm tells us is that each of a certain class of statements in ML is true. And to do so, it must presuppose we already understand the concept of truth in ML. But, of course, to suppose that we understand this is to make the whole effort at explaining truth in L quite unnecessary, since L is simply a part of ML. Thus the paradigm (1) is of no use in endowing 'true-in-L' with a tolerable degree of intelligibility unless we already understand 'true-in-ML'. And to assume that we do is simply to beg the question. Indeed, if, with Quine, we are willing to countenance 'true-in-ML' in MML then 'true-in-L' is easily definable in MML:

(6) (x) x is true-in-L if and only if x is a statement in L and x is true-in-ML.

Is there any way to gloss (1) that does not demand we already understand 'true-in-ML'? One attempt that might seem promising is:

(7) For all x, y and z, if x is a statement in L and y is the quote name in ML of x and if z is obtained by writing first y and then 'is true-in-L', then x and z are materially-equivalent-in-ML.

(This is a variation on a suggestion by J. F. Thomson in [9].) This alternative eliminates the use of 'true-in-ML', but only at the cost of allowing 'materially-equivalent-in-ML'. It might be thought that there is no gain, for after all material equivalence is defined in terms of truth. But actually we are a bit ahead since, though given 'true-in-L' and 'false-in-L' we can define 'materially-equivalent-in-L', the converse is not true. Nonetheless, if we assume the acceptability of

'materially-equivalent-in-ML' we have assumed enough to define 'true-in-L' allowing that we know some truth of ML. And this last bill is easy to fill since ML includes English as a part. So we have:

(8) (x) x is true-in-L if and only if x is a statement in L and x is materially-equivalent-in-ML to '1 = 1'.

The conclusion to be drawn is clear. The paradigm (1) is of no help in understanding the attribution of truth to any particular statement in L unless we already understand 'true-in-ML' or 'materially-equivalent-in-ML'. But if we allow either of these, the paradigm seems beside the point, since we have already accepted enough apparatus to define 'true-in-L'.

If what we have said so far is correct, then Quine's argument for the asymmetry between individual attribution of truth and individual attributions of analyticity evaporates. For just as 'true-in-L' can be rendered intelligible given 'true-in-ML' so 'analytic-in-L' can be defined, given 'analytic-in-ML'. Replacing 'true' by 'analytic' throughout in (6) will do nicely. Nor does glossing (1) in terms of material equivalence as in (7) help matters. The analogue of material equivalence in the theory of meaning is mutual entailment (or analyticity of the bi-conditional). And allowing the concept of mutual-entailment-in-ML, it is no trick to define analyticity-in-L. We have, as a direct analogue of (8):

(9) (x) x is analytic-in-L if and only if x is a statement in L and x and '1 = 1' entail one another in ML.

6. In the foregoing reflections we have been led to reject the claim that (1) is of any help in clarifying individual attributions of truth to statements. Now what of the remaining virtue claimed for (1)? This, it will be recalled, is that while not providing a definition, (1) at least gives us some way to tell whether a definition we might come up with is right or wrong. In particular, we are assured that any two acceptable definitions will select the same statements. It might be thought that this assurance is small solace without some clear idea of what we are saying when we call a statement true. But at least truth appears to retain some advantages over the concepts of the theory of meaning. The appearance is deceptive.

Let us imagine a semantic theorist who succeeds in explaining to our satisfaction the "English" binary connective '≡' for which there is no non-technical English equivalent. He might first explain 'analytic-in-English' say by recursively specifying which statements

are analytic. He could then go on to explain that a statement formed by writing any statement in English followed by '≣' followed by any statement in English is well formed and is true if and only if the same expression with 'if and only if' replacing '≣' is analytic-in-English. We do not assume that he has been resourceful enough to get us to buy 'analytic-in-L' for variable 'L'—only that he has clearly specified the extension of the predicate 'analytic-in-English'. He needn't follow this line, however. For present purposes we need only assume we understand '≣' as an English connective. Note that an analogous assumption about 'if and only if' has been made throughout our discussion of truth.

Now our resourceful theorist, having studied his Quine, might bemoan the fact that he can give no general definition of 'analytic-in-L' for variable 'L'. But, in spite of this he may claim that he can endow the expression with a high enough degree of intelligibility that we are not likely to be averse to using the idiom. He proceeds as follows:

First we must learn the object language L. Next we pool L with ME (viewed, now, as containing '≣'). This composite language will serve as ML. This done, we can offer the following paradigm:

(10) '———' is analytic-in-L if and only if (——— ≣ 1 = 1).

By way of explanation, he follows Quine, offering the remark we have quoted on page 388. As for the importance of his paradigm, again, he repeats Quine's argument, substituting 'analytic' for 'true' where appropriate.

Before our ingenious theorist has finished he will have met some strong protests. And the place where the shouting begins should be pretty clear. For in our assumption we granted '≣' as an English connective. But by the quick trick of pooling English and L he has begun to use '≣' as a connective in the composite tongue ML. What is more, he is using it in (10) between a statement of ML descended from L and one descended from English. Yet for *this* usage we have had no explanation. He might remedy this situation by using 'analytic-in-ML' and explaining (in MML) the use of '≣' in ML. The unhappiness of this course should by now be evident.

What is interesting about this little fable is that it finds a direct analogue in Quine's treatment of truth. Granting 'if and only if' in ME it becomes part of ML and is found in (1) between an expres-

sion deriving from ME and one deriving from L. Explaining this use of the locution is, presumably, one of the bothersome details we left to one side in pooling ME and L. But now how *can* it be explained except by recourse to 'true-in-ML'? (Lest it be cause for unwarranted optimism, let me observe that even if, as Quine maintains, we can give a pragmatically sound procedure for translating truth functional connectives from one language to another ([4], Sec. 13), this still gives no explanation for the use of these connectives (from either language) when they occur, as in (1), between a statement in one language and a statement in another.)

These reflections should give us a dim view of Quine's trick of pooling an object language with ME. But if we are unwilling to allow this move, then the remaining virtue claimed for (1) dissolves. For (1), or rather what we get from it by writing a statement of L for the blanks, is a statement in such a composite language. And, as Quine observes, unless we allow ourselves the expedient of pooling tongues, such substitutions in (1) result only in "a meaningless jumble of languages" ([3], p. 135).

7. There is, in Quine's writings, yet a third argument—or hint of an argument—for the superiority of the theory of reference over the theory of meaning: "In Tarski's technical construction," Quine writes, ". . . we have an explicit general routine for defining truth-in-L for individual languages L which conform to a certain standard pattern and are well specified in point of vocabulary" ([3], p. 138). Tarski's work, Quine claims, has provided explicit directions for defining truth-in-L for languages meeting certain general constraints. While it does not provide a general definition of 'true-in-L' for variable 'L', it does provide a way of constructing a variety of particular definitions of 'true-in-L' for languages of the appropriate sort. Here, then, we seem to have found some justification for preferring the theory of reference to the theory of meaning.

But let us attend more closely to the "explicit general routine" Tarski's work provides. Tarski does not formulate such a routine, though it is easy to see how it might be extracted from his work. Consider a simple example, a language containing finitely many one place atomic predicates whose sentences are formed using the devices of quantification theory. Basically, Tarski's technique is to specify the conditions under which an atomic sentential function is satisfied by an infinite sequence of objects, then to state how conditions of satisfaction are combined by the operators, quantifiers and

connectives available in the language ([8], Sec. 3). To begin a Tarski-type definition using English (or better, ME) as our meta-language, we must first list all the atomic predicates and find for each a translation into ME. Then the first clause of our definition will be a set of sentences of the form:

(11) A sequence s satisfies Px_i if and only if the i^{th} member of s is T_p.

where we replace 'Px_i' by the name (in ME) of the expression formed by appending the i^{th} variable to some predicate and replace 'T_p' by the translation of the predicate named into ME.

For Quine this is an unhappy beginning. In it we have made use of the concept of *translation*. The object language predicate and its ME translation are to be *synonymous*. So specification of the "explicit general routine" of which Quine speaks must make use of the concepts of the theory of meaning. Far from showing the theory of reference better off than the theory of meaning, the present line of defence uses a notion of the latter theory to clarify a notion of the former. Let us consider two or more languages to be the same *basic* language if they are identical in point of extra-logical predicates. If we want to use a single basic meta-language for all our definitions it is not easy to see how the present problem might be avoided.

At this juncture Quine might grant that Tarski's work provides no routine for defining truth-in-L in a single basic meta-language. But, he might continue, it does provide a routine for constructing a truth definition for each L (of the appropriate sort) in a different ML, viz. a meta-language containing L as a part. Here we need make no appeal to synonymy; the relation between the meta-linguistic expression replacing 'T_p' and the predicate named by the expression replacing 'P' in (11) is identity.

Yet Quine is still not out of the woods. For consider the second step in the procedure for constructing a truth definition of L, using as ML a language containing L as a part. We begin by identifying the connectives, quantifiers and operators of L. Then we construct, in ML, a definition of the form:

(12) A sequence s satisfies a sentential function f if and only if either
  (a) there is a sentential function f' such that f is the negation of f' and s does not satisfy f', or

(b) there are two sentential functions f' and f'' such that f is the disjunction of f' and f'', and either s satisfies f' or s satisfies f'', or

(c) there is a sentential function f' such that f is the universal quantification of f' under the nth variable and every sequence which differs from s in at most the nth place satisfies f',
etc.

But now notice that if the definition in ML is to provide any illumination on the idioms of truth and satisfaction we must be able to *identify* universal quantifications in L. Further, since the "explicit procedure" tells us to construct a definition of the form of (12) *in ML*, we must be able to *translate* such terms as the 'is' of identity and the universal quantifier 'every' into ML. Yet on Quine's view we can do neither. He maintains that in identifying and translating quantifiers, identity and other referential apparatus, no possible evidence can arbitrate between a variety of competing hypotheses.

> The categoricals depend for their truth on the objects . . . of which their component terms are true; and what those objects are is not uniquely determined by stimulus meanings. . . . Of what we think of as logic, the truth functional part is the only part the recognition of which, in a foreign language, we seem to be able to pin down to behavioral criteria. ([4], p. 61.)

So it appears Quine's third attempt to salvage the theory of reference runs afoul of his own arguments on the indeterminacy in radical translation of referential systems. The routine Tarski's work provides for constructing truth definitions sheds no light on the concept of truth in an exotic tongue unless we can identify the apparatus of reference in that tongue.

For those—and I am among them—who share Quine's view on meaning and on radical translation, these reflections point toward an uncomfortable conclusion. The theories of reference and meaning are beset with much the same problem. Reference is not rescued by Tarski's work. So if we are to adjure using the concepts of meaning we must, in good conscience, also abstain from the concepts of reference. If we are to respect our Quinian conscience we must abandon much philosophical thought about language and much of modern logic as well.

References

[1] Black, Max, "The Semantic Definition of Truth," in Max Black, *Language and Philosophy,* Ithaca, New York, 1949.
[2] Geach, P. T., "Designation and Truth," *Analysis,* v. 8, no. 6, 1948.
[3] Quine, W. V., "Notes on the Theory of Reference," in W. V. Quine, *From a Logical Point of View,* second edition, Cambridge, Mass., 1961.
[4] ———, *Word and Object,* Cambridge, Mass., 1960.
[5] ———, "Meaning and Translation," in J. A. Fodor and J. J. Katz, eds., *The Structure of Language: Readings in the Philosophy of Language,* Englewood Cliffs, N.J., 1964.
[6] ———, "Speaking of Objects," in *The Structure of Language, op. cit.*; also in W. V. Quine, *Ontological Relativity and Other Essays,* New York and London, 1969.
[7] ———, "Ontological Relativity: The Dewey Lectures 1968," *The Journal of Philosophy,* v. LXV, no. 7, 1968; also in *Ontological Relativity and Other Essays, op. cit.* Page references in the text are to *The Journal of Philosophy.*
[8] Tarski, A., "The Concept of Truth in Formalized Languages," in A. Tarski, *Logic, Semantics and Metamathematics,* Oxford, 1956.
[9] Thomson, J. F., "A Note on Truth," *Analysis,* vol. 9, March 1949.