SEMANTIC PRIMITIVES AND LEARNABILITY

Stephen LEEDS

One often hears it said that a language with infinitely many semantic primitives is unlearnable. This principle, which I shall call the Learnability Principle, is, if correct, of great importance in analyzing the logical forms of sentences in natural language: if an analysis, however plausible it may be on other grounds, turns out to ascribe to English an infinity of semantic primitives, that analysis must be rejected. And indeed this has lately seemed to be the fate of several well-known analyses of logical form. In this paper, I want to point out an ambiguity in the Learnability Principle which has the effect of turning it into two principles; of these, I shall argue that one is not obviously correct; the other, although very plausible, does not seem to be particularly useful.

The distinctions I want to make could be made in the context of any of the theories which are sometimes called compositional semantics—e.g., those of Church, Davidson, or Montague; I shall for the most part orient my discussion toward Frege, leaving the mutatis mutandis to the reader. What is typical of compositional semantics is a pairing of each sentence with another entity—a truth-value, a set of sequences, an intension—which we may call the semantic value of that sentence. The pairing of sentence with semantic value is achieved by one or another kind of composition: certain sub-sentential expressions (usually words) are assigned semantic values directly; compositional processes then generate the semantic values of more complex expressions, and ultimately those of the sentences. In Frege's theory, expressions are assigned two semantic values—a sense and a denotation; there are two sets of compositional processes, one which derives the sense of a complex expression from the senses of its parts, one which derives the denotation of a complex expression from the denotations of its parts.

Implicit in the above is a distinction between what we may call *complex* and *simple* expressions. A complex expression is one whose semantic value is compositionally derived from the semantic values of its parts; a simple expression is one with a semantic value which is not so derived. Simple expressions may of course be notationally complex; they may even consist of parts which have semantic value — thus, the simple expression 'rattan' may be held to contain the parts 'rat' and 'tan'; what makes 'rattan' nonetheless simple is that its semantic value does not arise compositionally from those of 'rat' and 'tan.' [1]

What is it to understand an expression ? Not all writers on compositional semantics have addressed themselves to this question; of those who do, the most forthright is perhaps Frege, who held that to understand an expression was to grasp its sense. There is a strong suggestion in Frege that whoever understands the simple constituents of an expression will have no difficulty in understanding the expression itself: if I have grasped the senses of 'Cicero' and '——is happy' (and if I correctly analyze the syntax of 'Cicero is happy'), then I already have grasped, or perhaps can in some way easily construct, the sense of 'Cicero is happy.'

As for simple expressions, Frege offers a sufficient condition for grasping their sense which is considerably less metaphorical than the above: one grasps the senses of at least some simple expressions by associating to them the correct definite description. Thus, to grasp the sense of 'Cicero', one associates with that name the definite description 'the Roman orator who denounced Cataline.' The associated definite description must

[1] It follows that the distinction between simple and complex is relative to theory: which compositional processes you discover will determine which expressions you call complex. I doubt if there is an absolute distinction: I do not imagine that there is a decisive reason to see 'impure' as derived from 'pure' or the converse, or to take both as simple. I should mention that my account leaves open the possibility that, in Frege's theory, one and the same expression might be complex from the point of view of the theory of senses, and simple from the point of view of the theory of denotations, or conversely. Other Fregean doctrines rule out this horrible possibility.

in turn be understood by grasping the senses of the words that appear in it; presumably the chain ends with simple expressions which are understood, not by associating definite descriptions, but in some other way.

In the case of proper names—the most frequently discussed kind of simple expression—there is a clear sense in which we may say that they are learned piecemeal. There is no calculation method that will enable you to predict the sense of a proper name you haven't seen before, say 'Cato'; you must ask people who Cato was, read a Roman history, or the like; each proper name presents a separate problem in learning. This fact about proper names has often been thought to apply to simple expressions generally, but it is important to see that nothing in Frege's semantic theory entails that this must be the case. Suppose with Frege that to learn the sense of a simple expression it is sufficient to associate with it a certain definite description (which itself is understood—I shall henceforth take this condition as read.) This leaves open the possibility for there to be mnemonic devices or rules of thumb which help us associate expression with description. One might imagine such devices existing on a large scale: there might be an infinite class of simple expressions C, and an effective map f (effective in the recursion-theoretic sense), such that f pairs each member of C with the correct, 'sense-giving', definite description. The existence of such a map would in no way threaten the status of the members of C as simple: so long as we do not analyze the members of C as deriving their senses or denotations from those of their parts they will continue to count as simple. I will argue that such devices in fact exist.

Let us consider quotations. Each quotation denotes its interior; thus '*rx px*' denotes *rx px*. (I use italicizing as a quotation-forming device in the metalanguage, to avoid an excess of quotation-marks). Should quotations be treated as complex, or as simple ? To treat them as complex presents considerable difficulties within Frege's theory: for example, if we take each symbol within quotation-marks (counting the blank space as a symbol) as denoting itself, then we are led to take the quotation-marks as denoting the function f such that, e.g.,

$f(r,x,—,p,x) = rx\ px$. f then is a function of variable degree; there is, however, no place for such functions in Frege's system. This difficulty does not show that one *cannot* take quotations as complex; it does make it attractive to treat them as simple; the class C of quotations will then be an infinite class of simple expressions.

How then can we represent quotations as learnable ? To each quotation, say '*rx px*' we may associate the definite description: *The expression here preceded by a colon: rx px.* Or, if one prefers a more structural description, one may associate with '*rx px*' the description: *The expression consisting of an are followed by an ex followed by an empty space followed by a pee followed by an ex.* In either case, each quotation is understood by associating the correct definite description; notice that the association is achieved by an effective map. Perhaps the neatest way to represent the learning of quotations—a method, it must be admitted, somewhat alien to Frege's point of view—is to see us as learning to understand quotations in virtue of coming to master an infinite recursive set of axioms, typified by:

1) 'rx px' is the expression enclosed by quotation-marks in sentence 1)

Other analyses of quotations are perhaps possible; my present point is that there are no obvious considerations which exclude the above analysis—in particular, no reason to suppose that our language contains only finitely many simple expressions. We come now to be the promised ambiguity in the Learnability Principle. What is a semantic primitive; in particular, are quotations semantic primitives ? One might choose to interpret 'semantic primitive' to mean 'simple expression'; read this way, the Learnability Principle is not obviously correct: to show it correct, one would at the very least have to show that the above analysis of quotations was unacceptable. A second version of the Learnability Principle arises if we read 'semantic primitive' as applying only to those expressions whose meanings cannot be learned by mastering a general rule which applies to many expressions; this version would not count quotations as semantic primitives; it would let the above analysis stand. I have no quarrel with this second version of the principle.

A version of the quotations example which is instructive to look at is an extension of written English containing infinitely many new constants which denote themselves. For example, any small object that can be pasted onto a page might be taken as such a constant; such a language might well turn out to be infinite [2]. Here there is no difficulty about how we learn the new constants: whenever we see one of them, we are by that very fact in an excellent position to identify its referent; Frege would then say that we had associated a sense with the name [3]. Here I think one might say that these constants were not semantic primitives, on the grounds that we have a general method for assigning a meaning to each constant; the second version of the Learnability Principle is thus not in conflict with this example. As for the first version, one might try to save it by refusing to see our new constants as simple, indeed as constants at all: one might read

2) @ is a blot of ink

as having the logical form of

3) This is a blot of ink: @ [4]

[2] This sort of language is not new. Jonathan Swift imagined something of the sort. Use of ordinals to name themselves is a familiar device in set theory.

[3] We could say that the sense of each of the new constants was given by the phrase 'The object here pasted on the page.' But Frege would have seen no need to say this: if we can identity the object, we have given its name a sense.

[4] But see how little point there is to this. If we introduced only 25 constants of the new kind, no one would object to 2) as it stands. How many constants would we have to introduce to compel reading 2) as 3) ? Why is the correct answer 'infinitely many' ?—I am no more capable of assimilating 10^{15} bits of information than I am capable of assimilating infinitely many. Perhaps, if my mental capacity is 10^{7} bits and yours is 10^{7+2} then a

The question is, of course, whether there is any reason to try to save the first version of the principle; I shall say something about this later.

I come now to an important example which I will not try to fit into a specifically Fregean picture. Consider a type-theoretic language L. The following will be subscripts: o, also $\alpha \rightarrow \beta$, for any subscripts α and β. Thus $o \rightarrow o$, $o \rightarrow (o \rightarrow o)$ etc. are subscripts. The sentences of L will resemble those of quantification theory, except that every variable and constant will bear a subscript. L will contain the sign '='; we will not use any other predicates. The interpretation of L is as follows: Individuals, say people, will be said to be entities of type o; entities of type $\alpha \rightarrow \beta$ will be functions that carry entities of type α to entities of type β. (Thus, entities of type $o \rightarrow o$ will be functions from individuals to individuals.) Variables subscripted α will range over entities of type α; constants subscripted α will denote entities of type α. For any $c_{\alpha\rightarrow\beta}$, d_{α}, $c_{\alpha\rightarrow\beta}(d_{\alpha})$ will denote the value of the denotation of $c_{\alpha\rightarrow\beta}$ when the denotation of d_{α} is taken as argument; similarly for variables.

Pick any compositional semantics you like for first-order quantification theory; you will find its extension to L obvious and inevitable. For definiteness, suppose we have in mind (as implicitly above) a denotational semantics. Let us now suppose that the constant Cicero_{o} denotes the man Cicero, and let us also suppose that for each α our ontology contains an entity of type $\alpha \rightarrow o$ which maps every entity of type α to Cicero. For each α let us denote this 'constant function of type $\alpha \rightarrow o$' by $\text{Cicero}_{\alpha\rightarrow o}$. Consider now the

language with $10^7 + 2$ constants of this kind is acceptable for you, but not for me ? Surely mental capacity is a red herring here: a language with these new constants will be learnable in the same way, regardless of how many constants it contains.

infinite set of constants '$\text{Cicero}_{o \to o}$', '$\text{Cicero}_{(o \to o) \to o}$', in general '$\text{Cicero}_{\alpha \to o}$'. I claim that it is at the least enormously difficult, and perhaps impossible, to represent these expressions as other than simple. For example, if we analyze '$\text{Cicero}_{o \to o}$' into denoting parts, what could we possibly take as the denotation of 'Cicero' (unsubscripted) ? The only plausible candidate is a function which ranges over all types, or perhaps over all the entities of every type; in either case, the denotation of 'Cicero' is not to be found in the ontology we assigned to L. It seems reasonable to conclude that L contains denumerably many simple expressions.

Shall we conclude then that L is unlearnable; that three generations of type theorists have not understood the language in which they worked ? More plausible, we may say that to come to understand the infinite set of Cicero-constants, one merely commits to memory a very simple, recursive set of axioms, namely those given by the following schema:

$$4)\ (\forall x_{\alpha})(\text{Cicero}_{\alpha \to o}(x_{\alpha}) = \text{Cicero}_{o})\ [5]$$

Again we are in conflict with the first version of the Learnability Principle; the second, as usual, is consistent with our analysis.

I have been arguing that a set of expressions which the smoothest and most natural semantic theory will treat as simple may nonetheless be, considered merely as notations, complex and structured; that we can sometimes take advantage of

[5] I am assuming here that the variables and quantifiers used in this schema are already understood. I do not have an account of how it is that we come to understand *them*—as obviously we do—but what seems to me difficult here is the question, what constitutes our understanding of any quantifier—not specifically the type-theoretic quantifiers. If our understanding of the first-order quantifier could be said to consist in our mastery of certain rules of inference, I would cheerfully transpose those rules into the type-theoretic context.

such notational complexity to devise a method for coming to understand every member of the set. The method will vary from case to case: in the first two examples I took our understanding to consist in our ability to recognize the referent, given the name; in the third example, our understanding consisted in our possession of a recursive set of axioms. All three examples seem natural, even obvious, yet the conclusion to which they point — that the first version of the Learnability Principle is false — has been resisted; we must now ask why.

Explicit arguments for the first version of the Learnability Principle are not easy to find in the literature — the best-known discussions seem to confuse the two versions [6]. Here is, however, a line of thought which is sometimes implicit in such discussions: A semantics may be thought of as a map which associates to each expression of our language its semantic value; to learn our language, it is thought, is in some way to grasp this map, whence the map must be effective. This requirement of effectiveness is then thought to place substantial constraints on what a semantics for a learnable language could look like: in particular, only a semantics which reveals a finite number of simple expressions will meet (or, perhaps, can be counted on to meet) the effectiveness requirement [7].

The idea that to understand our language one must master

[6] For example, Donald DAVIDSON, 'Theories of Meaning and Learnable Language,' in *Logic, Methodology and Philosophy of Science* (Yehoshua Bar-Hillel, ed.), Amsterdam, 1965.

[7] Sometimes this argument seems to be abetted by a dreadful pun on the words 'recursive,' 'recursion.' The correct observation that a semantic theory shows how the semantic values of complex expressions is generated by recursion (i.e. by iterative processes) from those of their parts is confused with the idea that a semantic theory shows the semantic values of complex expressions to be generated recursively (i.e. effectively) — the semantic values of simple expressions are not generated by recursion and so cannot be given by an effective process. I am about to deny that the notion of effectiveness has any application when we are discussing maps from expressions to semantic values, but even if it did make sense to speak of effectiveness here, we would want to bear in mind that a) not all recursive functions can only be defined by recursion, b) not all functions defined by any old sort of recursion will be recursive.

an effective procedure for associating expression with semantic value is not an easy one to take literally once we take into account the sorts of things semantic values are. Consider, for example, extensionalist semantics. Here the semantic values of 'Cicero,' 'Cicero's father,' 'Cicero's father's father,' etc. are various longdeparted Tuscans. Do we have here an effective procedure for associating expression with semantic value ? What could this possibly mean — that you and I know how to locate or recognize Cicero's father ? Or consider intensionalist semantics, say of the possible-world variety [8]. Here the semantic value of 'the largest city ever' is a function which picks out of each possible world a certain city (or else is undefined); is there a clear sense in which our understanding of the expression can be said to consist in our grasp of this function — a function whose value we cannot give for this world, let alone for others ? In possible-world semantics the semantic value of Fermat's Last Theorem is either the set of all worlds or the empty set — could we be said to understand Fermat's Last Theorem in virtue of having associated with it one of these sets (only we don't know which one), or in virtue of having an effective procedure that will allow us to do so ?

I have no doubt that there is something we do which can be somewhat metaphorically described as grasping the semantic value of an expression. But the metaphor must be cashed: in some cases to grasp the semantic value is to be able to recognize the referent, in others to know how to use the expression in inferences. Here the notion of effectivenss is indeed of considerable importance: if to understand quotations is to be able to recognize what each quotation denotes, we must have an

[8] The idea that to understand an expression is mentally to attach to it the right semantic value is more placsible in intensional than in extensional semantics: one surely does not understand a sentence by associating with it its truth value. Curiously, the impulse to say that we understand expressions by associating *something* with them has been so strong as to lead some extensionalists to invent a new kind of entity, a *truth-condition,* which we are said to grasp in understanding a sentence. Truth-conditions are, it seems to me good old-fashioned propositions; this view is intensionalism in disguise.

effective means of recognizing the denotations; if to understand the Cicero-constants is to have mastered axioms or rules of inference for them, these axioms and rules must be decidable. But such effective means of recognition, or rules of inference, can be given as easily for infinite classes of simple expressions as for complex expressions; to suppose that the map given by a semantic theory is in any clear sense effective, or that we can see, merely by examining the structure of that map, whether the language is learnable, is look for effectiveness in the wrong place.

By way of conclusion, we have been concerned almost exclusively with the first version of the Learnability Principle. The second version, which holds only that if we can understand an infinity of expressions, we must have some general procedure for coming to understand them, is surely correct. It does not seem, however, to be of much use. The examples I presented of analysis which conflict with the first but not the second version of the principle were not chosen quite at random: they are a fair sample, I think, of the sorts of analyses which the Learnability Principle has been thought to refute. Whether, once we have disentangled the two versions, there remain any interesting analyses which the second version of the Principle rules out, is a question I shall not try to answer — my own suspicion is that the answer is no.

University of Colorado, Stephen Leeds