THE LANGUAGE OF SET THEORY

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy in the Graduate
School of The Ohio State University

By
John Eric Nolt, B.A., M.A.

* * * * *

The Ohio State University

Reading Committee:

George F. Schumm
Charles F. Kielkopf
Steven E. Boër
Robert Kraut

Approved By

Adviser
Department of Philosophy
ACKNOWLEDGMENTS

I am indebted to William Lycan, Steven Boër, Charles Kielkopf, Robert Kraut, William C. Powell, Ronald Laymon, B. Chandrasekaran, and Steven Humphrey for stimulating dialogue, both in and out of the classroom. To my adviser, George Schumm, go special thanks, not only for philosophical conversation, but also for a great number of perceptive suggestions and corrections in all stages of the preparation of this manuscript. His guidance, criticism, and encouragement were invaluable, and his meticulous editing and typing of the final draft did much to enhance its coherence and readability. Finally, I wish to thank my wife, Jo, for careful typing of the preliminary draft and for heroic patience with a husband who often seemed to be away in another world.
VITA

Dec. 9, 1950 . . . . . Born - Canton, Ohio

1972 . . . . . . . . . B.A., Social Sciences, The Ohio State University Columbus, Ohio

1974-1978 . . . . . Teaching Associate The Ohio State University Columbus, Ohio

1976 . . . . . . . . . M.A., Philosophy, The Ohio State University Columbus, Ohio

FIELDS OF STUDY

Major Field: Philosophy of Mathematics


Studies in Logic. Professors George F. Schumm and Charles F. Kielkopf

Studies in the Philosophy of Language. Professors William G. Lycan and Steven E. Boër

Studies in Metaphysics. Professor Robert Kraut

Studies in Aesthetics. Professor James P. Scanlan
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vita</td>
<td>iii</td>
</tr>
<tr>
<td>Chapter</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>(1.1) The Trouble with Sets</td>
<td></td>
</tr>
<tr>
<td>(1.2) Notation and Other Conventions</td>
<td></td>
</tr>
<tr>
<td>2. Abstraction and Modality</td>
<td>23</td>
</tr>
<tr>
<td>3. Tarski Hierarchies</td>
<td>53</td>
</tr>
<tr>
<td>4. ZF</td>
<td>75</td>
</tr>
<tr>
<td>5. NF and Kripke's Theory of Truth</td>
<td>104</td>
</tr>
<tr>
<td>6. Set-Theoretic Knowledge</td>
<td>117</td>
</tr>
<tr>
<td>7. Some &quot;Dogmas&quot; Revisited</td>
<td>143</td>
</tr>
<tr>
<td>8. Set Theory and Metaphysics</td>
<td>171</td>
</tr>
<tr>
<td>Notes</td>
<td>185</td>
</tr>
<tr>
<td>Bibliography</td>
<td>194</td>
</tr>
</tbody>
</table>
1.1 The Trouble with Sets

Textbooks on set theory usually begin with a statement to this effect: "A set is a group or collection of things." Shortly thereafter we are introduced to the null set, a set which has no members. If we take the initial characterization seriously, then we must think of the null set as a group or collection of no things. But the notion of such an object seems paradoxical, even self-contradictory. For as we ordinarily understand these terms, there can be no group or collection unless there is at least one (and, in the most common usage, more than one) thing of which it is a group or collection. A group or collection of no things—if we could make sense of this idea—would seem to be just nothing at all. But, according to set theory, this "collection" exists. Therefore, what the textbooks tell us, though perhaps pedagogically expedient, is misleading: the mathematician's concept of a set is quite unlike ordinary notions of groups or collections.

The extent of this difference is often surprising. We do not, for example, ordinarily talk about collections
or groups of one thing. Yet if we adopt this way of speaking for the moment, it seems most natural to say that the collection whose sole member is this marble ought simply to be this marble. What else could it be? And yet in set theory this "collection"—the unit set of this marble—is not the marble, but something distinct from it.¹

The existence of a null set and the distinction between unit sets and their members is the source of another difference between collections and sets. For from the null set and succeeding unit sets there arises an endless series of so-called "pure" sets—sets whose only members are other sets, which themselves have only sets as members, and so on. Such, for example, is the series

$$\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \ldots$$

Indeed, in all but the very weakest set theories, series of this sort can be extended transfinitely to nondenumerable length. Our ordinary conception of a collection involves nothing like this at all. There are no pure collections; every collection (including collections of collections) is composed ultimately of noncollections. And this means that the number of groups or collections is dependent in a very direct way upon the number of individuals in the universe. If the universe consisted of three objects, a, b and c, for example, the total number of collections would be four: the collections consisting of a and b; a and c; b and c;
and a, b and c. Moreover, if there were no such basic objects, there would be no collections. According to set theory, on the other hand, the number of "collections" in the universe is already uncountably infinite before we introduce even one basic individual.

These differences point to a more fundamental one. Insofar as they are determinate, at least, the identity conditions for collections differ from those of sets. Sets are the same just in case their members are the same, but collections may be the same, even though they are collections of different things. For example, the sets \{a,b,c\} and \{a,\{b,c\}\} are not identical, since they have different members. The first has three members; the second, two. But our everyday understanding, though imprecise, seems to draw no distinction between the collection consisting of a, b and c and the collection consisting of a and the collection of b and c. Accordingly, the identity of collections is not simply a function of their "members." It hinges only upon the individuals from which they are ultimately constructed, and not at all upon the identity of intermediate collections produced in the construction process.

A collection, therefore, is rather more like what Goodman calls a "sum individual" than like a set. A sum individual is a whole composed of parts. There are no "unit" sum individuals distinct from their parts, and there
is no "null" individual—that is, an individual without parts. (Each object is a part of itself.) These are features which are shared by the ordinary notion of a collection. It would be a mistake, however, to identify this ordinary concept with Goodman's precise technical notion, for ordinary usage is imprecise and ambiguous, and it does not conform in all ways to Goodman's axioms.

The important point is that sets are not the sorts of things we usually think of as groups or collections. The set-concept is something unique; it has no counterpart in ordinary language. It is important to make this plain at the outset of any attempt at a conceptual clarification of set theory, for it is easy to be lulled into thinking that sets as classically conceived are perfectly ordinary objects, like collections of china or groups of flowers.

No one, so far as I am aware, finds collections or groups in the ordinary sense especially puzzling, and no reasonable person doubts their existence. This page is a collection of molecules, and its existence is at least as certain and unproblematic as that of the molecules themselves. More importantly, like all collections of physical things, this collection itself is a physical thing, composed of matter and situated in space and time.

Sets, however, are not easily regarded as physical. This is obvious in the case of the pure sets. (If the null set is a physical object, then where is it and why has it
not been observed?) But failure to appreciate the differences between sets and collections has made it less obvious in the case of sets of physical things. Metaphysicians sometimes offhandedly identify physical collections with sets. One occasionally hears the suggestion, for example, that ordinary physical objects are just sets of more fundamental things—molecules, point-events, or the like. But though every collection of physical things is a physical thing, not every set is a physical thing. Consider the set of all physical things. Since, on standard set theories at least, no set is a member of itself, this set cannot be physical. One could, of course, deny that such a set exists; but if one believes in sets in the first place, reasons for doing so are difficult to muster.

Nor is the identification of other sets with physical objects without difficulty. Again, identity conditions are the crux of the matter. The identity condition for physical collections is the same as that for other physical things: sameness of spatiotemporal location. (Location, of course, does not imply localization. The constellation Orion is a collection of stars situated at widely-varying distances from the earth, but its identity as a collection is precisely delimited by specification of the spatiotemporal locations of its components. It is an object no different in principle from more familiar objects, such as this page, whose positions exhibit much less "scatter.")
But the identity conditions of sets are unlike those for physical objects, and as a result assertions of identity between sets and physical objects pose serious conceptual difficulties. This emerges in a striking way if we try to identify sets of physical objects with physical collections having the same members. If a set of physical objects is the collection of those objects, then the location of the set is just the locations of its members. This page, for example, considered as the set of its constituent molecules, is located at all and only those places its molecules are. But if this is true in general—that is, if the location of every set of physical things is just the locations of its members—then the unit set of this page, itself a "collection" of physical things (in this case just one), has the same spatiotemporal location as the page. So by the criterion of identity for physical objects, it is this page. But by the criterion of identity for sets it cannot be, since the page and its unit set do not have the same members.

This contradiction can be avoided only by the admission that, while this page is a physical object, its unit set is not and therefore has no spatiotemporal location. But this maneuver is ad hoc, because it divides "collections" of physical objects into two groups—those which are themselves physical objects, and those which are not (the unit sets)—without independent motivation.
Moreover, it fails to solve the problem. Consider the following two sets: (i) the set of all molecules of this page, and (ii) the set with two members, one of which is the set of all molecules comprising the top half of this page, and the other of which is the set of all molecules comprising the bottom half. Both (i) and (ii) are physical "collections"--(ii) because it is the collection of two such collections--and neither is a unit set. But (i) and (ii) occupy the same space at the same time and hence are identical. According to set theory, however, they are distinct.

We could retreat still further. If we adopt an atomistic metaphysics, we can claim that the only sets which are physical collections are sets of atoms. (Again without independent motivation, we have barred more sets from the realm of the physical.) This position escapes contradiction, but it still has some unpleasant consequences. We must admit, for example, that though this table top and each of the four legs attached to it is a physical object, the "collection" of all five is not the table, as we would expect--indeed, not a physical object at all.

Thus at the cost of atomism and a great deal of unnaturalness, we can construe some sets as physical objects. But what is to be gained? Since many other sets are evidently nonphysical anyway, the result is a rather gerrymandered bifurcation of sets into those which inhabit the physical universe and those which do not. Since our
metaphysician is stuck with nonphysical sets anyway, his ontology will be neater if he regards them all as nonphysical and observes carefully the distinction between collections and sets.

But nonphysical sets pose problems of their own. If we cannot see them or hear them or detect them with geiger counters, how can we know about them? Among those who believe in nonphysical sets, two hypotheses are widely held at present: (1) the pragmatic view that we know about sets because we find it useful to "posit" them in the process of unifying and systematizing scientific knowledge,³ and (ii) the "sixth sense" hypothesis—the idea that we somehow "perceive" or "grasp" sets by a faculty of mathematical intuition.⁴

The pragmatic view is initially attractive, but it generates a very difficult problem. We know about sets because positing them is useful, but why is it useful? Unless there is some tie—causal or structural or otherwise—between sets and the physical universe, their existence or lack of it could not make the slightest difference to us. There must be such a tie, if the pragmatic view is correct, but its nature remains a metaphysical puzzle to which the pragmatist has no solution.

The "sixth sense" hypothesis, on the other hand, raises some hope of explaining why sets are useful, but as an account of set-theoretic knowledge it is blatantly ad hoc.
The sole explanatory role of the extraordinary mental faculty it posits is to account for this knowledge; and, aside from introspective testimony by its proponents, no independent evidence has ever been offered in its favor.  

Epistemological puzzles are not our only worries if we take nonphysical sets seriously. There are conceptual and metaphysical ones as well. What does it mean to say that a nonphysical object exists? We know what it means for a physical object to exist: it has a more or less definite location in space-time and various causal interactions with other physical things. But to say that the null set exists is not to say that it has spatiotemporal position; not to say that it has causal relations with anything; indeed, not to say that it makes any palpable difference to the physical universe at all. So what do we mean when we say that it exists? We certainly do not mean it is there. By hypothesis, it is not anywhere.  

Our inability to give an informative explanation of what we mean when we say that a set exists might be regarded as a mere curiosity and annoyance if in practice we could tell which sets exist and which do not. Then even if we could not explain our meaning, we could at least be sure that our assertions meant something definite. But that is not the case. In recent years the proliferation of theories of classes and sets, many with claims to interesting applications, has been explosive. The existence claims of
many of these theories are incommensurable. And though
mathematicians have for the present settled by and large on
a core of axioms (those which comprise Zermelo-Fraenkel set
theory), there is little conviction, even among mathemati-
cians, that these axioms embody the one true theory (i.e.,
that they tell us which sets actually exist, while all
conflicting theories are false).

Among those who do believe this, the principle
justification is that these axioms produce the most elegant
foundation for classical mathematics. This is regarded as
inductive evidence for their truth and hence for the exist-
tence of the sets of which they speak. The strength of this
evidence is open to serious question, but to examine the
matter thoroughly would require a lengthy digression into
the topic of scientific explanation. Instead, let us turn
the matter on its head and grant hypothetically that the ZF
axioms are known to be true. Even supposing this much, it
still can be seen that we cannot tell in practice whether
sets whose existence is not determined by these axioms
exist or not.

The source of this insight is the wealth of indepen-
dence results which have been obtained in set theory in the
past several decades. These results show that even after
the axioms of ZF have been assumed, there remain countless
ways of consistently adding existence postulates, many of
which are incompatible. If we can tell in practice which
sets not determined by the ZF axioms exist, then we ought to be able to decide among these additional existence assumptions. In fact, however, there is good reason to believe that we cannot.

The original ZF axioms, whose truth we are for the moment granting, were settled upon because they yielded the most satisfying proofs of classical mathematical results. They were chosen for their ability to account for already-existing theorems. But now that all the classical results have been accounted for, power in proving existing mathematical results can no longer be used as a criterion to decide the truth of further axioms. The unresolved questions of set existence which face us at present are not questions which have arisen because of our need to provide foundations for results we already have. Rather, they are questions which, resolved one way, yield an entirely new array of mathematical results and, resolved differently, yield different but perhaps equally intriguing results. And because these existence questions are independent of our present axioms--and hence of present mathematical theorems--no present results can dictate choice among them.

This situation has often been likened to that in geometry after the emergence of noneuclidean systems. Just as no geometry seems to be the one true geometry (though some are more useful than others for describing the physical manifold in which we find ourselves), so too
there seems to be no uniquely true set theory. Some are more interesting than others; some are more useful or more elegant; but the lesson of the independence results and the consequent proliferation of alternatives seems to be that none is the one true theory.

Our conclusions, then, are two. First, we are not able to give a clear, easily intelligible explanation of what it means to say that a set exists; second, even granting that the mathematical community has arrived at a consensus on some of the axioms of set theory, it is unlikely that this consensus can be extended much further. Moreover, it is doubtful that we should grant this much. The proliferation of alternatives to ZF raises the still more difficult question of whether there are multitudes of set-universes, only some favored ones, or none at all. There seem to be no principled grounds for accepting any of these choices, and we may seriously wonder whether the notion of existence as applied to platonic sets has any determinate meaning at all.

These considerations, together with the epistemological ones raised earlier, suggest that the platonic view of sets is seriously amiss. It is the purpose of this study to present an alternative view.

Whatever sets are (if, indeed, they are anything), it is clear that there is a peculiarly intimate connection between at least some of them and unary formulas—that is,
formulas in one free variable. For it is by means of such expressions that we refer to sets, and it is by their means that sets are defined. The set of all blue things, for example, which is customarily denoted by the expression '{x | x is blue}', is defined by the unary formula 'x is blue'. The formula can be used to define the set because its satisfaction conditions mirror the set's membership conditions; an object satisfies the formula if and only if it is a member of the set. Now since the only properties of sets that are of any practical interest are their membership conditions, and since these membership conditions are precisely reflected in its defining formula or formulas, it appears that we can dispense with talk of sets altogether in favor of talk about their defining formulas, and lose nothing in the bargain. Instead of understanding set theory as a theory of nonphysical "collections," we could then regard it simply as an oblique way of talking about language. Such a shift, could it be accomplished, might carry us a long way toward solutions to the problems recently discussed. This, roughly, is the alternative view we shall suggest.7

On the face of it, the idea has several intriguing features. By treating membership as satisfaction, we merge the so-called "logical" and "semantic" paradoxes, and (it is to be hoped) generate insight into both. Russell's paradox about the set of all sets which are not members of themselves, for example, becomes just an alternative formulation of the
Grelling paradox, which, on one version, concerns the unary formula 'x does not satisfy itself'. Traditionally, these paradoxes have been regarded as unrelated and their obvious similarity has been viewed as mere coincidence. But our proposal would explain this similarity by showing that the two paradoxes are at bottom identical.

Furthermore, if set theory is a theory of the satisfaction conditions of unary formulas (i.e., a kind of semantic theory), then it ought to be part of a more general theory of truth—a theory which deals not only with unary formulas, but with formulas of all sorts. It is noteworthy that the inconsistent comprehension principle of naive set theory,

(1) $\exists x \forall y (y \in x \leftrightarrow A(y))$, for any formula $A(y)$,

when written in instantiated form,

(2) $y \in \{x \mid A(x)\} \leftrightarrow A(y),$

bears a strong formal resemblance to Tarski's principle

(3) $A$ is true iff $\text{tr}(A)$

(where $\text{tr}(A)$ is a translation of $A$ into English). This resemblance is heightened by our proposal, on which (2) is to be read as

(4) $A(x)$ is satisfied by $y$ iff $\text{tr}(A(y))$. 
Taken together, (3) and (4) suggest the following generalization:

\[(5) \quad A(x_1, \ldots, x_n) \text{ is true (of } y_1, \ldots, y_n, \text{ respectively) iff } tr(A(y_1, \ldots, y_n)), \text{ where} \]
\[A(x_1, \ldots, x_n) \text{ is any formula in } n \text{ free variables.} \]

Here 'true (of)' functions as a variably polyadic predicate, satisfiable by both single objects and n-tuples. If we represent it by 'T', then we may write "'Pain hurts'T", meaning that 'Pain hurts' is true, or "'x is cold'T(Pluto)", meaning that 'x is cold' is true of (i.e., satisfied by) Pluto, or "'x is taller than y'T(Omar,Lea)", meaning that 'x is taller than y' is true respectively of Omar and Lea, and so on.

Now (5), like (1), is inconsistent. A version of the Grelling or Russell paradox follows from it almost immediately. Moreover, the principles expressed in (5) are involved in the liar paradox and a host of others. Yet even though (5) is inconsistent, something like it must be true, since many of the inferences licensed by its biconditional are valid. We correctly reason, for example, that if 'x = x' is true of 2, then 2 = 2, and vice versa.

In the following chapters, we shall regard set theory as a fragment of a theory of truth—or, more precisely, we shall regard various set theories as fragments of truth theories for various languages. We shall explore the ways
in which principles like (1)-(5) are related. And, we shall see that when set theory is viewed as semantics, many of the metaphysical, epistemological and conceptual problems which plague it become soluble. Before turning to the details of our proposal, however, a few words about methodology are in order.

This document presents an analysis of set-theoretic language. Opinions differ widely on what an analysis is or should; so rather than engaging in a complicated and tangential discussion of the nature of analysis, I would simply like to explain how I will be using the word.

An analysis, as I understand it, is a proposal for understanding one way of speaking in terms of another. Analysis is called for when a customary way of speaking is found to be philosophically troublesome. Under such circumstances, we may find it illuminating to "paraphrase" or "translate" this way of speaking into another which accomplishes the same pragmatic purposes while avoiding the philosophical difficulties of the first. Analysis does not involve giving the meaning of a way of speaking, nor need it involve synonymy. But a good analysis must do at least two things: (i) the "translation" it gives must preserve the utility of the original way of speaking, and (ii) the "translation" must be less subject to philosophical problems than the original.
The problematic way of speaking which interests us is the language of set theory. Set theory, as we noted, is useful in a variety of mathematical, linguistic and philosophical enterprises. But it is also philosophically troublesome in ways we have mentioned. Our task, then, is to show that set talk can be "translated" into talk about unary formulas in a way that avoids or at least illuminates these difficulties while retaining the utility of set theory for mathematics, semantics and other applications.

The analysis will involve no claim that when a mathematician or semanticist speaks of sets, what that person is really talking about, without being aware of it, is unary formulas. The platonist intends to be talking about sets--genuine ghostly sets--and it would be perverse to insist that he cannot say what he intends. We may quite reasonably insist that what he is talking about does not exist and that he is mistaken, but not that he means something utterly different from what he thinks he means.

What we hope to show, then, is only that the practical goals which set theory, understood platonistically, allows mathematicians and others to attain can be attained as well by philosophically less objectionable ways of speaking. But of what consequence is this? At most it shows that one seductive argument for the existence of sets--namely, that we cannot do mathematics and semantics without speaking of them--is wrong.
Yet that in itself is significant. Few arguments for the existence of sets are nearly as convincing as this one, and lack of convincing evidence for an existence claim is as good a reason as any to doubt it.

Our work begins in Chapter 2 with a deepening and broadening of the analysis proposed here. We suggest that talk about abstract entities (particularly sets and formula types) should be regarded as talk about concrete things by the method of identifying indiscernibles. This method alone, however, gives an inadequate analysis and must be supplemented by a modal interpretation of quantifiers.

Chapter 3 considers the relation of Tarski-style hierarchies of languages to theories of types and to the "universes" of various set theories. We show that some weak forms of set theory may be regarded as quantifying over the unary formulas of such hierarchies.

In Chapter 4 we define a more complex transfinite hierarchy of languages and show that on the assumption of ZF's consistency it is a model of ZF. This indicates that the analysis can be extended to powerful set theories.

Chapter 5 discusses Kripke's theory of truth as a way of dealing with the satisfaction relation that is more streamlined and elegant than Tarski hierarchies. We suggest that Kripke's theory has close ties with Quine's set theory New Foundations (NF), and we speculate that it may prove useful in the search for a consistency proof for
that system.

The last three chapters return to philosophical themes. Chapter 6 takes-up the question of set-theoretic knowledge and argues that our analysis is epistemologically superior to traditional platonistic views. From this discussion it emerges that set theory, on our analysis, is analytic or true by convention, as the positivists urged. Chapter 7 recounts the reasons for the rejection of the positivistic philosophy of mathematics by Quine and others, and defends the notion of truth by convention against Quinean objections. Finally, Chapter 8 examines the metaphysical implications of the analysis and assays its impact on such fields as semantics and modal logic.

1.2 Notation and Other Conventions

Throughout this work we shall use ~, →, and ∃ for negation, material implication, and the existential quantifier, respectively, with ∧, ∨, ↔, and ∀ defined in terms of them in the usual manner. Capital letters from the beginning of the alphabet will stand for formulas and other expressions, lower case letters from the beginning of the alphabet for objects and sometimes constants, and lower case letters from the end of the alphabet for variables. The letters 'F', 'G', 'H', and 'P' will stand for predicates, and 's' and 't' for terms. Lower case Greek letters denote ordinals, ω being the set of nonnegative integers.
The vocabulary of a language of level \( \beta \) consists of
the operators \( \neg \), \( \rightarrow \), and \( \exists \), brackets ( and ),
the variables \( v^n \) (\( n \in \omega \), \( 0 \leq \alpha \leq \beta \)),
at least one \( i \)-ary predicate \( P^i \) for
some \( i \in \omega \), and possibly some special symbols to be dis­
cussed later. We shall have no need for primitive singular
terms. In referring to variables, we frequently omit
superscripts and subscripts.

To define a language \( L \) of level \( \beta \) we first specify a
set of atomic formulas. The formulas of \( L \) are then defined
by the following recursion:

1. Each atomic formula is a formula of \( L \).
2. If \( A \) and \( B \) are formulas of \( L \), then \( \neg A \) and \( (A \rightarrow B) \)
   are formulas of \( L \).
3. If \( A \) is a formula of \( L \) and \( x \) is a variable of the
   vocabulary of \( L \), then \( \exists x A \) is a formula of \( L \).

A classical language is a language \( L \) of level \( 0 \) such that
\( A \) is an atomic formula of \( L \) if and only if \( A \) is \( P^i x_1 \ldots x_i \),
where \( P^i \) is an \( i \)-ary predicate of the vocabulary of \( L \) and
\( x_1, \ldots, x_i \) are variables of \( L \). The scope of a quantifier
\( \exists x \) is the smallest formula immediately following that quan­
tifier. An occurrence of a variable \( x \) is bound if it is in
the scope of a quantifier \( \exists x \); otherwise, the occurrence is
free. A variable is free in a formula \( A \) if at least one
occurrence in \( A \) is free. A sentence is a formula in which
there are no free variables. A unary formula is a formula
with exactly one free variable.

A variable assignment for a language \( L \) of level \( \alpha \) on domains \( D_\beta (\beta \leq \alpha) \) is a function which assigns to each variable \( x_\alpha^n \) of \( L \) a member of \( D_\alpha \).\(^8\) An interpretation \( I \) of \( L \) on domains \( D_\beta \) is a (possibly partial) function which assigns the values \( T \) or \( F \) to pairs consisting of a formula and a variable assignment.\(^9\)

A standard predicate assignment for a classical first-order language \( L \) on domain \( D \) is a function which assigns to each \( i \)-ary predicate of \( L \) a member of \( D^1 \). A classical interpretation for a classical language \( L \) on domain \( D \) is induced by a standard predicate assignment \( p \) as follows: for all formulas \( A \) of \( L \) and all variable assignments \( f \) on \( D \),

\[
(1) \text{ if } A \text{ is } F^i x_1 \ldots x_i, \text{ then } I(A,f) = T \text{ or } F \text{ according as } <f(x_1),\ldots,f(x_i)> \in p(F^i) \text{ or not;}
\]

\[
(2) \text{ if } A \text{ is } \neg B, \text{ then } I(A,f) = T \text{ or } F \text{ according as } I(B,f) = F \text{ or } T;
\]

\[
(3) \text{ if } A \text{ is } (B \lor C), \text{ then } I(A,f) = T \text{ or } F \text{ according as } I(B,f) = F \text{ or } I(C,f) = T, \text{ or not; and}
\]

\[
(4) \text{ if } A \text{ is } \exists x B, \text{ then } I(A,f) = T \text{ or } F \text{ according as } I(B,f^{d/x}) = T \text{ for some } d \in D, \text{ where } f^{d/x} \text{ is the function just like } f \text{ except that it assigns } d \text{ to } x, \text{ or not.}
\]

In Chapter 2, when we speak without qualification of "languages" and "interpretations," it is to be understood that we
are referring to classical first-order languages and to classical interpretations. Beginning in Chapter 3 we shall deal with more sophisticated languages and interpretations.

The notation $\models A(x_1, \ldots, x_n)$ is used to designate formulas and indicate facts about their free variables. Without stipulation to the contrary, it is assumed that $x_1, \ldots, x_n$ include all of $A$'s free variables, and perhaps some other variables as well.

$\models_I A(x_1, \ldots, x_n)[a_1, \ldots, a_n]$ abbreviates $I(A, f) = T$ for all variable assignments $f$ such that $f(x_i) = a_i (1 \leq i \leq n)$. Likewise, $\models_I A(x_1, \ldots, x_n)[a_1, \ldots, a_n]$ abbreviates $I(A, f) = F$ for all $f$ such that $f(x_i) = a_i (1 \leq i \leq n)$.

We often abbreviate lists of the form $[a_1, \ldots, a_n]$ as $[a]$, so we might write $\models_I A(x)[a]$ in place of 
$\models_I A(x_1, \ldots, x_n)[a_1, \ldots, a_n]$. Occasionally, we shall omit mention of free variables altogether and write just $\models_I A[a]$. And sometimes we write simply $\models_I A$, meaning that $A$ is true on all variable assignments. Boldface Gothic letters are used to shorten other expressions as well. For example, we might write $\models_I \delta(a)$ in place of $\models_I \delta(a_1), \ldots, \delta(a_n)$ or $\models_I \langle c \rangle$ in place of $\langle c_1 \rangle, \ldots, \langle c_n \rangle$, but this practice should cause no confusion.

If $A(x)$ is an open formula with $x$ free, we say that $d \in D$ satisfies $A(x)$ on interpretation $I$ over domain $D$ if and only if $\models_I A(x)[d]$. 
CHAPTER 2
ABSTRACTION AND MODALITY

Our proposal is to analyze talk of sets as talk of unary formulas, but as it stands this proposal is ambiguous. 'Unary formulas' could mean either 'unary formula types' or 'unary formula tokens'. Formula types are abstract objects, the forms which different concrete bits of language "share" or "have in common." Formula tokens, on the other hand, are concrete inscriptions and utterances, things evidently compatible with a physicalistic ontology.

Analysis of sets in terms of formula types would result in some conceptual and ontological simplification. Instead of both the membership and satisfaction relations, we would need only the latter; and instead of two sorts of abstract objects, we would require only one. But it would not help with the conceptual problems attached to the notion of abstract or platonic existence; and it would not advance our understanding of set-theoretic knowledge, since knowledge of platonic types is as problematic as knowledge of platonic sets.

If we could show that set theory can be reconstrued as talk about formula tokens, on the other hand, then not
only would we reap the benefits of a reduction of sets to types, but also these additional problems would be well on their way to solution. Talk about abstract objects would be eliminated altogether, and problems about our knowledge of sets would reduce to the familiar problems about our knowledge of the physical world. Analysis in terms of formula tokens is therefore clearly preferable.

Seeming talk about abstract objects can sometimes be analyzed as talk about concrete things, as Quine notes in "Logic and the Reification of Universals" (1953):

It may happen that a theory dealing with nothing but concrete individuals can conveniently be reconstrued as treating of universals, by the method of identifying indiscernibles. Thus, consider a theory of bodies compared in point of length. The values of the bound variables are physical objects, and the only predicate is 'L', where 'Lxy' means 'x is longer than y'. Now where '~Lxy . ~Lyx', anything that can be truly said of x within this theory holds equally for y and vice versa. Hence it is convenient to treat '~Lxy . ~Lyx' as 'x = y'. Such identification amounts to reconstruing the values of our variables as universals, namely, lengths, instead of physical objects (p. 117).

The quantifiers of Quine's theory actually range over physical bodies, but when 'x = y' is substituted for '~Lxy . ~Lyx' as a matter of definition, the theory seems to be quantifying, not over bodies, but over their lengths. However, as Quine is quick to point out:

This method of abstracting universals is quite reconcilable with nominalism, the
philosophy according to which there are really no universals at all. For the universals may be regarded as entering here merely as a manner of speaking—through the metaphorical use of the identity sign for what is really not identity but sameness of length... In abstracting universals by identification of indiscernibles, we do no more than rephrase the same old system of particulars (pp. 117-18).

It will be useful to examine the method of identifying indiscernibles more closely.

DEF 2.1: Objects a and b are discernible by a formula $A$ of language $L$ under interpretation $I$ if there are $c$ in the domain of $I$ such that it is not the case that $\models_I A[a,c] \neq \models_I A[b,c]$.

DEF 2.2: Objects a and b are discernible in a language $L$ under interpretation $I$ if a and b are discernible by some formula of $L$ under I.

DEF 2.3: Objects a and b are discernible by a predicate $P$ of an interpreted language $L$ if a and b are discernible by some atomic formula of $L$ containing $P$.

Objects not discernible by a formula, in a language, or by a predicate are said to be indiscernible in these respects. It is easy to show by induction on the length of formulas that if two objects are indiscernible by each predicate of an interpreted language, then they are indiscernible in
The following result expresses one of the most useful properties of indiscernibles:

**THM 2.1**: Let \( a = a_1, \ldots, a_n \) and \( b = b_1, \ldots, b_n \). If \( a_i \) is indiscernible from \( b_i \) in \( L \) under \( I \) for each \( 1 \leq i \leq n \), then for all formulas \( A \) of \( L \), \( \models_I A[a] \iff \models_I A[b] \).

**Proof**: Suppose \( a_i \) is indiscernible from \( b_i \) in \( L \) under \( I \) for each \( 1 \leq i \leq n \), and let \( A \) be any formula of \( L \). Now \( a_i \) and \( b_i \) are indiscernible by \( A \) for each \( i \). So by repeated application of **DEF 2.1**, \( \models_I A[a] \iff \models_I A[b_1, a_2, \ldots, a_n] \iff \models_I A[b] \).

If two objects are indiscernible in \( L \), then everything which can be said in \( L \) of one can also be said of the other. With respect to the predicative resources of \( L \), they appear to be the same object.

In Quine's example, objects of the same length are indiscernible. His language ignores or abstracts away from all properties of objects other than the relational ones of being longer or shorter, simply by lacking the predicates which express these properties. As a result, the language is insensitive to differences among objects of the same length. All objects of the same length have all of their expressible properties in common. With respect to anything which can be said of them, they are identical. Sameness of length is thus what might be regarded as the standard of
identity relative to Quine's language.

This notion of language-relative standard of identity can be made precise. A binary predicate expresses the standard of identity operative within a given language if, whenever a pair of objects satisfies that predicate, everything which can truly be said in the language about one can also truly be said about the other. This is the case if the predicate in question obeys the two classical axioms of identity, the law of reflexivity and the law of substitutivity. We shall call such a predicate an identity predicate.

DEF 2.4: A formula $A(x,y)$ with just $x$ and $y$ free is an identity predicate for a language $L$ under interpretation $I$ if

1. $I \models \forall x A(x,x)$, and
2. $I \models \forall x \forall y ((B \land A(x,y)) \rightarrow B(\frac{y}{x}))$,

where $B$ is any formula of $L$ not containing $y$ and $B(\frac{y}{x})$ is the result of replacing one or more occurrences of $x$ in $B$ by $y$.

Quine's predicate '$\sim Lxy \cdot \sim Lyx$' is an identity predicate, for it clearly satisfies (i) and can be shown by induction on the length of formulas to satisfy (ii).

By adopting an identity predicate whose interpretation is weaker than absolute identity (i.e., which is true of distinct pairs of individuals), we produce a language which
treats distinct individuals as if they were the same. If absolute identity is expressible in our language, however, abstraction will not be possible, since any two objects will be discernible. More generally, if our language contains any identity predicate at all, it cannot also contain a weaker one (one which discerns fewer pairs of objects); within a single language, there can be only one standard of identity.

**THM 2.2:** If \( A(x,y) \) and \( B(x,y) \) are identity predicates for a language \( L \) under interpretation \( I \), then

\[
\models_I \forall x \forall y (A(x,y) \leftrightarrow B(x,y)).
\]

**Proof:** By Clause (i) of DEF 2.3, \( \models_I \forall x B(x,x) \), and by Clause (ii) \( \models_I \forall x \forall y ((B(x,x) \land A(x,y)) \rightarrow B(x,y)) \). Hence \( \models_I \forall x \forall y (A(x,y) \leftrightarrow B(x,y)) \). The converse is similar.

In view of THM 2.2, we may speak loosely of the identity predicate for an interpreted language, since all identity predicates for a given language are equivalent.

The significant property of identity predicates is that a pair of objects satisfies the identity predicate for a language if and only if the objects are indiscernible in that language.

**THM 2.3:** Let \( A(x,y) \) be an identity predicate for \( L \) under \( I \). Then objects \( a \) and \( b \) are indiscernible in \( L \) under \( I \) if and only if \( \models_I A(x,y)[a,b] \).
Proof: Let \( a \) and \( b \) be indiscernible in \( L \) under \( I \). By DEF 2.4(i), \( \models_I A(x,y)[a,a] \). So by THM 2.1, \( \models_I A(x,y)[a,b] \).

Conversely, suppose that \( \models_I A(x,y)[a,b] \). Then by THM 2.1, \( \models_I A(x,y)[b,a] \). So by DEF 2.4(ii), for any formula \( B \) of \( L \) and \( c \) in the domain of \( I \), \( \models_I B[a,c] \iff \models_I B[b,c] \). Hence by DEF 2.1 and DEF 2.2, \( a \) and \( b \) are indiscernible in \( L \) under \( I \).

THM 2.3 is in effect Leibniz' Law, with both identity and indiscernibility relativized to interpreted languages. What counts as identical or indiscernible when identity and indiscernibility are conceived this way is a function of the predicative richness of one's language.

One can ask, once the concept of relative identity has been introduced, whether there is an absolute sense of identity at all, or whether all identity is relative in this way. Fortunately, this difficult question need not be answered here. It suffices for our purposes to note that languages can be ranked (partially ordered) according to their ability to discern objects. We shall call a language which fails to discern objects discernible in a richer superlanguage abstract relative to that superlanguage. Thus Quine's truncated idiom is abstract relative to the language of natural science. The question of whether in languages still richer than those to which we are accustomed further levels of discernibility emerge, though fascinating, is irrelevant to our investigation. Our interest lies in
the other direction—in increasing abstraction—not in increasing discernibility.

We wish to clarify talk about abstract objects—sets and formula types. And the recent digression on indiscernibility suggests that such talk may best be understood as talk in abstract languages about concrete things. Specifically, it suggests that talk about expression tokens, talk about expression types, and talk about sets take place in languages of increasing abstraction. That is, talk about types is talk about tokens in which we abstract away from or ignore those properties of tokens (such as spatial location, size, physical composition, etc.) which distinguish them from others of the same type. Tokens of the same type thus become indiscernible, so that our language, while still quantifying over tokens, appears to be talking about types instead (just as Quine's language, while still quantifying over physical objects, appears to be talking about their lengths). Set-theoretic language emerges, then, as a further level of abstraction—the level attained by ignoring all properties of tokens other than their satisfaction conditions (which are now thought of as membership conditions). All tokens with the same satisfaction conditions become indiscernible; and our identity predicate takes on the meaning 'are satisfied by the same things' or 'have the same members', the latter of which is the appropriate criterion of identity for sets.
Thus, for example, when we say that the null set is not a member of itself, we are not making a statement about some dark and numinous "aggregate of no things." Rather, we are saying something about open sentence tokens—specifically, about tokens used to define the null set. These are tokens of the form $x \neq x$, and their equivalents, and what we are saying about them is that they fail to satisfy themselves. Each statement ostensibly about sets or expression types is likewise reinterpretable as a statement in a relatively abstract language quantifying over open sentence tokens. The abstract seems to have been reduced to the concrete.

Besides those already mentioned, this reduction has a number of additional points in its favor. It predicts, for example, that set-theoretic language, being abstract in the manner just described, makes no use of predicates capable of discerning objects with the same satisfaction conditions—a prediction which is confirmed by set-theoretic practice. Except in philosophical contexts, we never apply predicates other than $\epsilon$, and those definable in terms of it, to sets. We neither affirm nor deny, for example, that predicates of size, color, mass, etc., are true of sets. Where set theory is actually put to use, such predicates are not employed. And when in philosophical contexts we do raise questions about the applicability of such predicates to sets, a host of metaphysical problems arises immediately—a good indication that something is seriously amiss. Of these
problems we shall have more to say shortly.

Moreover, the reduction illuminates the difficulties of identifying sets with physical objects (discusses in Chapter 1) is a very revealing way. Since the language in which we talk about sets is more abstract than the language in which we talk about physical objects, it has a correspondingly more liberal identity predicate. The difference in language—not any natural metaphysical difference—is what accounts for the difference in identity conditions between sets and physical things. Now any attempt to identify sets and physical things results in a mixing and confusion of the two languages. Importing the identity predicate for sets into talk of physical objects requires that we render all physical objects other than formula tokens indiscernible, since all of them have the same satisfaction conditions (they are satisfied by nothing). On the other hand, importing the identity predicate for physical objects (sameness of spatiotemporal location) into the language of sets renders equivalent formulas discernible and hence destroys the abstraction—the very feature which makes the language a language of sets. In neither case can we apply the identity predicate of the one language to the objects of the other. One cannot speak of both sets and physical objects from within the same language, and one cannot apply the same identity predicate to them. The difference between sets and physical objects emerges, then, not as an
ontological difference, but as a difference in the expressiveness of the languages used to speak of each. These remarks apply mutatis mutandis to types as well.

Despite these attractive features, the analysis as it stands has a fatal flaw. The world contains only finitely many expression tokens, but there are infinitely many sets and expression types. There is no way to generate this infinity of abstract entities by identifying indiscernible tokens. Identification of indiscernibles can only reduce, not increase, the number of objects we seem to be talking about. If we quantify only over existing tokens, the resulting theories will appear to assert the existence of only finitely many sets and types. But this is totally unacceptable. Even the most elementary mathematics cannot be developed in a finite set theory, and any syntactic theory which tried to make do with finitely many expression types would be severely crippled. Quantification over tokens renders many—indeed, infinitely many—sets and types ultimately "irreducible." The following sentence, for example:

(1) ∃x(x consists of 10^{100} symbols)

is true when the quantifier ranges over expression types but false if we regard it as ranging over tokens. There are no tokens that big. Statements of set theory likewise get the wrong truth-values. Suppose, for example, that as
a result of some catastrophe all open sentence tokens satisfied by nothing were destroyed. Then the set-theoretic statement

\[(2) \exists x \forall y \sim y \in x,\]

which ostensibly asserts the existence of the null set, would (reading \(\in\) as 'satisfies') be false. Thus set-theoretic statements and statements about types would vary in truth-value with the creation and destruction of tokens. And never would all of them have the right truth-values, since never would infinitely many tokens be constructed. There simply are not enough expression tokens to make the reduction work.

But though sufficiently many tokens do not actually exist, sufficiently many are possible. The reduction could be made to work if we let our quantifiers range, not over actual tokens, but over possible ones as well.\(^1\) Though (1) and (2) are false with their quantifiers limited to actual tokens, they both become true when the domain is extended to possible ones. Other statements about sets and types likewise regain their appropriate truth-values. Thus, though we cannot reduce the abstract to the actual concrete, we can reduce it to the possible concrete spoken of in an abstract idiom. Our analysis requires both abstraction and modality.
The expansion of our domain from actual to possible tokens renders a reduction feasible, but it remains to be seen whether it is still desirable. Those with Quinean sensibilities are likely to think not. Sets and types are well-behaved; their identity conditions are precise and thoroughly understood. But the identity conditions of possibilia seem shifty and ill-defined; talk about possibilia lacks precision and generates puzzles. Better, Quine would say, to leave talk of types unanalyzed than to analyze it as talk of possibilia.

But Quinean objections to talk of possibilia lose their force when that talk proceeds in an abstract language. For within such a language, the precise identity conditions of specific tokens do not matter; the language is not "fine-grained" enough to discern individual tokens anyway. Given a token of a particular type, for example, the type language lacks the ability to make distinctions between it and any other token of that type. The result is that Quinean perplexities do not arise. Are there two or \( N_{16} \) possible expression tokens in the doorway? The question cannot even be raised; our language lacks the predicates necessary to formulate it. We do not have the predicate 'is in the doorway', since it discerns possible tokens of the same type; and we cannot even speak of numbers of individual tokens, since that, too, requires their discernibility.
The abstraction weeds out the predicates in terms of which Quinean puzzles can be formulated. Lest this should seem to be an arbitrary and illegitimate evasion, we should hasten to add that at bottom it is no different from the way in which the platonist himself escapes similar difficulties. To see this, let us turn the issue on its head: How many platonic expression types are there in the doorway? The platonist traditionally offers one of two responses to such questions. The first is to declare the question nonsense—to say, for example, that it commits a category mistake. Types, a proponent of this response would hold, are not the sorts of things of which it can meaningfully be asserted or denied that they are in the doorway. The second answer is that there are no types in the doorway, because types are not in space at all.

The first response avoids trouble by declaring that certain predicates cannot meaningfully be employed in the context of talk about types. Having noted that application of spatial predicates, color predicates, and so on, to types would create embarrassing philosophical difficulties, the proponent of this view systematically prohibits use of these predicates in discourse about types. In other words, whenever he is talking about types, he ascends to a more abstract idiom. Hence the platonist of this stripe avoids Quinean difficulties in the same way we do—by rendering them inexpressible. Notice, too, that the predicates he
bans are precisely the ones absent from our abstract language--namely, those predicates capable of discerning tokens of the same type: predicates of size, color, location, physical composition, etc.

The second response, though superficially unrelated to the first, actually employs a similar strategy. This response involves the metaphysical claim that types are not in space or that they lack spatial properties. This claim, as it stands, is badly in need of clarification. But one thing it surely entails is that expression types do not satisfy spatial predicates. And this consequence alone suffices to resolve the Quinean question; for it entails that the number of types satisfying the predicate 'is in the doorway' is zero. Instead of banning unwanted predicates, this response neutralizes their harmful effects by declaring that their application to types consistently yields, not meaninglessness, but falsity. The troublesome predicates remain in the language but fail to do any work--and hence any damage--in application to types.

The corresponding move in our abstract language quantifying over possibleia would be to admit all predicates but to declare that the ones which generate problems (those which discern tokens of the same type) be considered to have null extensions in the context of discussion ostensibly about types. Tokens of the same type would thereby cease to be discernible in those contexts, and the net effect
would be the same as if we had ascended to an abstract language (though in fact what we have done is to remain within our richer language and alter the semantics of certain predicates in the context of purported talk about types).

Thus the platonist who answers queries about the location of sets and types by replying either that such queries are nonsense or that sets and types are not in space is just obliquely declaring his decision not to apply spatial predicates to platonic things. Our reply to similar queries about possible tokens, though more explicit, is essentially the same. If the platonist's response is adequate, then ours cannot be less so.

An important point which emerges from these considerations is that the language of platonic types and our abstract language quantifying over possible tokens are in nearly every respect the same. The similarity extends even to their identity predicates. In both cases the criterion of identity is what might be called syntactic similarity. Two tokens are indiscernible in the abstract language (and hence satisfy its identity predicate) if they have all syntactic properties in common (i.e., if they are of the same type). But this is also the criterion of identity employed in the platonic language (though here it is regarded as absolute identity). The only significant difference between the two languages is that we regard the variables of the latter as ranging
over actual abstract objects and the variables of the
former as ranging over possible concrete tokens.

Given the extent of the correspondence between the
two languages, and especially the fact that they employ
the same standard of identity, an important question arises:
What sense is there to the claim that we are actually quan-
tifying over possible tokens? The same question can be
raised when we make the further abstraction to sets. Is
what we are doing really any different from quantifying over
types and sets from the start?

There is a difference—and an important one—
especially in view of the epistemological question raised
earlier. Abstract objects, if we take them simply as given,
are peculiar things, thought to be causally unrelated to us
and disjoint form our spatiotemporal world. Our knowledge
of them seems inexplicable. But the values of variables of
our abstract language are not like that; they are, many of
them, concrete actual things which we utter or write and
with which we are intimately familiar. The set theorist who
believes himself to be quantifying over platonic sets cannot
point to anything in the world about which he is talking.
On the other hand, if we construe him as talking abstractly
about possible tokens, then we can actually exhibit some of
the objects of his theorizing. They are the formulas, the
open sentences, which he writes on a blackboard or prints
in a book—indeed, the very open sentences he uses to
"define" his sets. And, though some of what he is talking about is merely possible, and hence more puzzling, seeing that at least some of the time he is talking about bits of this world may help us to understand how he seems to make reference to otherworldly things. If he is really talking about platonic objects, then perhaps we can do no better in explaining his knowledge than to point out that he finds belief in them profitable or to claim that he has a special faculty of mathematical intuition. But if, at least some of the time, he is talking about fragments of the here and now, then maybe we can do better. We shall certainly be at an advantage in explaining his knowledge of those "sets" and "types" for which he has produced corresponding tokens; he has actually bumped up against what he is talking about. His knowledge of the others—those "sets" and "types" corresponding to merely possible tokens—though more perplexing, is surely not inexplicable.

It is not inexplicable because knowledge of this sort is fact—indeed, commonplace fact. While it remains highly dubious that we know anything categorical and non-tautologous about platonic objects, it is obvious that we have a great deal of substantive knowledge about nonactual expression tokens. We know, for example, that English has possible expression tokens containing $10^{100}$ symbols, and we know that every possible expression token of a classical first-order languages contains but finitely many symbols—
the list could go on ad nauseum. In Chapter 6 we shall see that this knowledge is explainable in terms of the familiar epistemic processes of deduction and inductive generalization. Thus even the modalized version of our analysis retains an epistemological advantage over the platonic view.

There are other advantages as well. The platonist who offers a pragmatic or wholistic account of our knowledge of abstract things is hard-pressed to explain how the entities he posits have any relation to practical affairs. On our account the explanation is straightforward. What we perceive through the distorting spectacles of abstract language as sets or types are in fact fragments of language—predicates and open sentences. And we can rather obviously use such bits of language to categorize, to count, to calculate, to compare—indeed, to do all the things for which set theory is useful. Moreover, the fact that we can do these things is no metaphysical mystery; many of the language-fragments in question are right at our fingertips—literally.

Yet in spite of these advantages, our analysis might appear to be a step backward ontologically. Platonic objects, though immaterial, are at least assumed to be actual; but many of the tokens in the range of our variables are not. Thus, while the platonist has only actual objects in his ontology, we seem forced to concede commitment to the existence of merely possible ones as well.
Terminological care is essential in evaluating this sort of objection. Traditionally, it is true, the quantifier \( \exists \) is read 'there exists', and its use commits us to the existence of the objects in its domain. But this reading is no longer appropriate when the domain includes nonactual objects. For, on the most common usage, 'exists' means 'is actual'—and we hardly intend to assert the actuality of nonactual objects. For domains containing mere possibilias, a more satisfactory reading is 'there could be' or 'there is possible', which avoids the misleading connotation of actuality.

Reluctance to accept quantification over mere possibilias stems in part, I think, from failure to appreciate this point. We habitually assume that quantification over something commits us to its existence. Since 'existence' connotes 'actuality', quantification over nonactual objects seems self-contradictory, and it makes us uneasy. When we read \( \exists \) as 'there is possible', we are less inclined to feel this uneasiness, since it becomes obvious that our commitment is not to the actuality of possibilias, but only to their being possible. Thus, for example, where the platonist asserts that there actually exist sentence types containing \( 10^{100} \) symbols, we say (within our abstract language) that there could be sentence tokens with \( 10^{100} \) symbols or that tokens with \( 10^{100} \) symbols are possible.
When the alternative are seen in this light, it is not at all clear that the platonic commitment is preferable. As we noted earlier, abstraction eliminates the sorts of modal paradoxes usually associated with quantification over *possibilia*. And, if the issue is one of the intelligibility of the notion of possibility itself, then we would do well to ask about the intelligibility of the notion of actual platonic existence as well, for the very meaningfulness of this notion, as we saw in Chapter 1, is open to serious question.

In particular, we raised three problems for the notion of platonic existence: (i) that we seem to be unable to give an illuminating explanation of what it means to say that a set exists, (ii) that even assuming the existence of some sets is generally agreed upon, there seems to be no systematic way of extending the agreement, and (iii) that the proliferation of alternative set theories casts doubt on the idea that there is a unique set-theoretic universe—and hence raises the seemingly unanswerable question of how many there really are.

The notion of existence employed in set theory is so poorly understood that there is little to lose in replacing it with the notion of possibility. And there is much to be gained. Unlike the concept of platonic existence, the notion of possibility needed for our analysis can be explained in a fairly clear and understandable way. Let
us ask, for example, what it means to say that tokens with $10^{100}$ symbols are possible. As a start, we should note that what counts as a possible token is language-relative. The domain of possible tokens is different for speakers of Chinese than it is for speakers of English. When we talk about possible expressions tokens, we must have some background language or languages in mind; otherwise, every possible object would be a possible token, since any possible object could be employed as an expression token of some possible language. Therefore, in quantifying over possible tokens, we always limit our domain to the possible tokens of some familiar language or languages. In asking what it means for tokens with $10^{100}$ symbols to be possible, then, we must have such a language or languages in mind. Let us suppose this language is English.

Now, very roughly, what it means to say that English tokens of a given description are possible is that we could, given sufficient time and paper, actually construct such tokens—not that we physically could (for perhaps tokens with huge numbers of symbols are physically impossible), nor yet that we logically could (since infinite tokens are logically possible but not—according to most grammarians, at least—possible in English), but that we could according to the grammatical rules of English. The sort of possibility with which we are concerned, then, is possibility with respect to grammatical rules. A token of a given description
is possible in English if tokens of that description are grammatical in English.

The same should apply to formal languages. But when a token satisfying a condition \( A(x) \) (e.g., 'x contains \( 10^{100} \) symbols') grammatical in a formal language? The grammatical rules of a formal language include its formation and interpretation rules. What is or is not grammatical depends entirely on what is stipulated in these rules and on nothing else. Classically, the formation and interpretation rules of a language are regarded as statements about formula types. They indicate which types exist and what their truth conditions are. Formula tokens satisfying the condition \( A(x) \) are regarded as grammatical if the rules imply the existence of a formula type satisfying \( A(x) \), i.e., if the statement

\[
(3) \exists x A(x)
\]

is deducible from the rules. (We assume that the variables of this statement range only over the formula types whose existence is stipulated by the rules.) Likewise, no formulas satisfying \( A(x) \) are grammatical if

\[
(4) \sim \exists x A(x)
\]

is so deducible. Such is the usual account of how the grammatical rules of a formal language determine what is and what is not grammatical in that language.
Our account is no different in substance. We agree that formula tokens satisfying $A(x)$ are grammatical in a language $L$ (and hence possible relative to $L$) if

\[(5) \ R \vdash \exists x A(x)\]

and not possible whenever

\[(6) \ R \vdash \neg \exists x A(x),\]

where $R$ is the set of $L$'s grammatical rules. We differ only in our understanding of what the rules and the statements deducible from them assert. Rather than regarding the rules as statements about the *existence* of formula types, we prefer to think of them as statements in an abstract language about the possibility of formula tokens. Thus (3), for example, should be understood as asserting that a formula token satisfying $A(x)$ is possible. Accordingly, a formula token satisfying $A(x)$ is possible relative to a language $L$ just in case the statement that such a token is possible is provable from the grammatical rules of $L$.\(^2\)

This equivalence suggests that we might analyze the notion of possibility with respect to a language in terms of the purely syntactic notion of provability, thus eliminating the former. But such an analysis would ultimately be circular. Provability itself is generally not taken as a primitive notion, but is defined in terms of the *existence* of a proof. Now, of course, we do not accept the platonic
notion of existence used in this definition, and will analyze the notion of the existence of a proof in terms of the possibility of a proof, i.e., the possibility of a certain arrangement of formula tokens of a language—and this creates a circle.

Possibility relative to a language must therefore remain a primitive notion for us. It is what is stipulated by grammatical rules and what we learn about by making deductions from these rules and applying them to concrete situations. Yet it is a primitive notion with an important relation to the concept of provability and hence in turn with relations to practical activities (e.g., the activity of performing deductions) which the concept of platonic existence, and particularly platonic set existence, lacks. This enables us to explain what it means for a token to be possible relative to a given language and to illustrate our explanation by giving sample deductions from the rules. It also helps to explain our knowledge of possible tokens, as we shall see in Chapter 6.

An important problem arises in connection with the grammatical rules for exotic languages (such as those with infinite expressions, which we shall encounter in Chapter 4) or incompletely defined languages (such as we shall encounter in Chapter 6 and later). The grammatical rules for such languages are incomplete, in the sense that for some formulas A(x) of the metalanguage in the which the grammatical rules
are expressed, neither (5) nor (6) is true. Incomplete rules leave a fringe of indeterminacy. For some conditions A(x) they fail to determine whether a token satisfying A(x) is possible or not.

When a set R of grammatical rules is incomplete, we may think of these rules as defining, not a single language, but a range of languages—specifically, those which could be fully defined by adding sufficiently many new rules (consistent with the original set) to attain completeness. (In some cases these new rules will be infinite in number and not recursively specifiable.) Now when (5) is true, where R is the original incomplete set of rules, tokens satisfying A(x) are possible in each of the languages in this range. When (6) is true, tokens satisfying A(x) are possible in none of them. And when neither is true, tokens satisfying A(x) are possible in some languages in the range, but not in others.

Every complete set of grammatical rules determines whether a token satisfying A(x) is possible for each condition A(x) of the metalanguage in which it is formulated. However, even a complete set of grammatical rules does not in general answer all the questions which can be asked about the language it defines. Formal formation rules, for example, do not ordinarily specify whether the tokens they define are to be written in ink or in crayon or to be constructed in some other way. The predicates 'written-in-ink',
'written-in-crayon', etc. are not employed in the language in which the formation rules are expressed. Thus they leave indeterminate the question of whether tokens written in ink are possible. We can, however, raise the question of the possibility of such tokens if we are speaking a richer language (say, English as a whole). Considered as statements of this richer language, the rules, which were complete relative to the poorer, more abstract sublanguage, now lose their completeness. From the point of view of this richer language, therefore, they are best viewed as defining a range of languages (those determined by the various possible ways of completing the rules) in the manner described earlier. Some of the languages of this range will have possible tokens written in ink; in others, such tokens will be prohibited. It thus emerges that whether we view a set of rules as defining a single language or a range of languages may depend upon the richness of the total descriptive apparatus in which the rules are formulated, as well as on whether the rules are complete with respect to the language containing only the predicates which occur in them.

This total descriptive apparatus can apparently be enriched without limit. Any set of rules which we shall ever have occasion to consider can thus be considered incomplete with respect to some superlanguage of the language in which the rules are expressed. As a result, it is never the case that all questions about possibility for
the tokens of a given language have determinate answers. For each set of rules which seems to define a single language, we can find a superlanguage from the point of view of which the rules are incomplete and hence define only a range of languages.

Normally, when dealing with formal languages, however, our metalanguage is, as we saw earlier, a fairly well-defined, relatively abstract language, from the point of view of which linguistic expressions appear as types. While from our richer English superlanguage, the rules formulated in this metalanguage define a range of languages, within the metalanguage itself, the rules, if complete, appear to define just one language; and every statement about the possible tokens of this language has a definite truth-value.

The fact that possibility for formula tokens (i.e., grammaticality) is relative to the language to which the tokens belong (as determined by the rules which define the language) provides an explanation for the proliferation of set theories—a bit of data which it is difficult to fit into the platonic scheme. If set theory is a theory about possible expression tokens, then the tokens in question must belong to some language or languages. And if different set theories say different things about possible tokens, then the obvious explanation is that these are tokens of different languages, each defined by a different set of grammatical
(in this case interpretation) rules.

As we shall soon discover, most set theories are best regarded as theories, not about tokens of a single language, but about tokens of an entire hierarchy of languages—a hierarchy akin to the hierarchies of object languages and metalanguages first described by Tarski. Different set theories, then, can be construed as theories of language hierarchies which differ in various ways. The possible formulas which appear abstractly as sets are formulas of hierarchical semantical theories.

To show that such a construal is possible, we must define the grammatical rules governing the hierarchies in question and then show that the tokens possible according to these rules model the set theory with which we are concerned. This we shall do for a very weak set theory (the Constructive Theory of Types) in Chapter 3, and for more complicated set theories (including $ZF$) in Chapters 4 and 5.

In all three chapters, we shall use some set theory in our metalanguage in the process of making these definitions and proofs. This procedure may appear illegitimate, for we seem to be assuming the existence of sets in our metatheory in order to prove that sets can be eliminated in favor of possible formula tokens. But this illegitimacy is only apparent. We initially regard the set theory of our metalanguage as "intuitive" or unanalyzed. That is, we shall assume that it is a part of our home language of which we
have some prior understanding, but which remains open to further clarification and analysis. After we have completed the analysis of the set theories of our various object languages, it will become clear that a similar analysis is applicable to our metatheory. Ultimately, therefore, we shall regard our metatheory as a theory of possible formula tokens.

In Chapters 3-5 the type-token distinction will play little role, and we shall usually speak simply of "formulas." In all such cases we shall mean 'possible formula tokens', in accordance with our analysis. Likewise, all talk of existence in these chapters is to be regarded as talk of possibility.
The language of standard set theories contains the two predicates $\in$ and $=$, in terms of which all other set-theoretic notions can be defined. The former is nearly always taken as primitive, but $=$ can be defined in terms of $\in$ thus:

**DEF 3.1:** $x = y$ for $\forall z (x \in z \iff y \in z)$.

If the variables range over sets, $=$ signifies membership in the same sets, but when sets are reduced to formulas, it signifies satisfaction of the same unary formulas. In this latter case, $x = y$ means that $x$ and $y$ are indiscernible (with respect to the unary formulas of the domain).

An important related notion is that of having the same members (on the classical reading) or extensional equivalence (on our analysis). This relation, represented by the predicate $\equiv$, is defined thus:

**DEF 3.2:** $x \equiv y$ for $\forall z (z \in x \iff z \in y)$.

The axiom of extensionality, which on the classical view states the identity conditions for sets, asserts that these
two relations are coextensive:

\[ \text{EXT: } \forall x \forall y (x = y \leftrightarrow x \equiv y). \]

Read classically, EXT asserts that sets are identical if and only if they have the same members. On our analysis it says that unary formulas are indiscernible if and only if they are extensionally equivalent. Thus it expresses the degree of abstractness of set-theoretic language. This language, it tells us, has no unary formulas capable of discerning extensionally equivalent formulas; and yet in it all nonequivalent unary formulas are discernible.

We shall call the classical first-order language whose sole primitive predicate is \( \epsilon \) the language of set theory or \( L^S \) for short. If we could state the satisfaction conditions for \( \epsilon \), we would have a complete interpretation of \( L^S \) from which we could calculate the truth-values of set-theoretic statements. Regrettably, however, all the straightforward ways of doing this lead to paradox. Since \( \epsilon \) itself means 'satisfies', the most obvious way to state its satisfaction conditions is as follows:

(1) For all unary formulas \( A, B \) of \( L^S \),

\[ \models x \epsilon y[A, B] \iff \models B[A]. \]

But this definition is circular; the satisfaction conditions of \( B \) are not defined until those of \( \epsilon \) are, and vice versa. The result is contradiction. To derive a version of the
Russell or Grelling paradox, let A and B both be \( \sim x \in x \).

What has gone awry? There are two ways of seeing the matter. From the set-theoretic side, it appears that we have stated a principle very closely related to the inconsistent comprehension principle of naive set theory:

\[
(2) \exists y \forall x (x \in y \leftrightarrow B), \text{ for any formula } B.
\]

But from the semantic side, what we have done is to try to make \( \epsilon \), a predicate of \( L^S \), function as the satisfaction predicate for \( L^S \)--which, as Tarski has shown, is impossible unless we allow truth-value gaps.\(^1\)

From the semantic point of view, the traditional way of avoiding paradoxes like those generated by (1) is to make a rigid distinction between object language and metalanguage. The satisfaction predicate for a language can occur only in its metalanguage, and the satisfaction predicate for its metalanguage occurs only in a metametalanguage, and so on. Thus is envisioned a sequence of languages, each expressing the satisfaction conditions for formulas of those "below" it. Such a sequence is a Tarski hierarchy.

The analogue in set theory of the stratification of languages induced by the object language-metalanguage distinction is the stratification of the set-theoretic universe into orders, levels or types. This is evident to some degree in almost all set theories, but most obvious in those based upon theories of types. In this chapter we shall
show how one simple type theory—the Constructive Theory of Types (CTT)—quantifies, on our analysis, over a Tarski hierarchy. We shall then discuss how various modifications of Tarski hierarchies provide "universes" for more sophisticated set theories. This coalescence of Tarski hierarchies and set-theoretic universes provides another example of the conceptual unity and simplicity which our analysis offers.

We begin with a brief and incomplete survey of the variety of Tarski hierarchies. We can think of every language as having its own Tarski hierarchy of which it is the lowest or base language. This hierarchy, however, is not unique; many different sorts of hierarchies can be built upon the same language. In view of our set-theoretic interests, we shall initially be concerned with hierarchies in which each succeeding language contains only a satisfaction predicate and quantifies only over unary formulas of lower languages. Later, in Chapter 8, we shall briefly consider Tarski hierarchies allowing quantification over formulas of all sorts and employing the variably polyadic predicate $T$, which was mentioned in Chapter 1.

Whether we allow only a satisfaction predicate for unary formulas or something more exotic, we face a variety of additional options. Each metalanguage of a hierarchy may be distinct from its object language(s), or each may contain its object language(s). Hierarchies of the first sort are discreet; those of the second sort, nested. Each language
of a nested hierarchy contains all previous languages, and the hierarchy itself can be viewed as a single superlanguage of all its member languages.

Within a given hierarchy some mechanism is needed for preventing confusion of language levels. There are at least two options. The first is to use a notationally different satisfaction predicate for each language of the hierarchy. This permits each language of the hierarchy to have the same style of variables, so that all the languages are interpreted over a common domain which contains the unary formulas of all languages in the hierarchy. Paradox is avoided by stipulating that each satisfaction predicate is false of all pairs of formulas of languages or level greater than or equal to the level of the language to which it belongs. The second method is to have a single satisfaction predicate for the entire hierarchy but to restrict the ranges of variables within a given language. This method requires a notationally distinct style of variables for each language. The first method we call predicate-typing; the second, variable-typing. Variable-typing is also the device used to prevent confusion of domains in theories of types. The two methods can also be used together, but their combination adds nothing new and is of little interest. Each alone is sufficient to prevent confusion of languages.

In a nested variable-typed hierarchy, special constraints on the occurrence of variables are often useful.
If such constraints are in force, the hierarchy is restricted; if not, it is unrestricted. A hierarchy is strongly restricted if the type of the variable to the right of the satisfaction predicate must be exactly one higher than the type of the variable to the left. A language is weakly restricted if the type of the variable to the right of the satisfaction predicate must be at least one higher than the type of the variable to the left.

In addition to these syntactic options, there are semantic ones. In a variable-typed hierarchy, variables of each type may range over all objects lower in the hierarchy or only over objects of the next lowest stratum. In the first case, the hierarchy is cumulative; in the second, noncumulative. It is most natural to interpret weakly restricted or unrestricted hierarchies cumulatively and strongly restricted hierarchies noncumulatively, but this is not necessary.

Finally, one can restrict the ranges of a hierarchy's variables in any of a number of other ways—for example, by specifying that only formulas of certain syntactic types are included. One particular restriction of this sort is of interest. Let us call a unary formula of a variable-typed hierarchy predicative if it contains no variables of a type higher than that of its free variable. A hierarchy, then, is predicative if its variables take only predicative formulas as values, and impredicative otherwise.
We shall now construct a nested, variable-typed, strongly restricted, noncumulative, predicative Tarski hierarchy. This structure is of interest because it models the Constructive Theory of Types and serves as a starting-point for the elaborations which will lead us to models of more sophisticated set theories.

To construct the hierarchy, we first define a base language $L_0$ and an arbitrary domain $D_0$ for the interpretation of $L_0$. We let $L_0$ be the classical first-order language whose only predicate is the one-place predicate $P$. The variables of $L_0$ are $v^0_1 (i \in \omega)$. For $D_0$ we can take any nonempty class of nonlinguistic objects—by which we mean objects not considered to have truth or satisfaction conditions. To interpret $P$, we assign it the null extension.

The hierarchy itself consists of the languages $L_n$ $(n \in \omega)$. The vocabulary for each $L_n$ $(n > 0)$ consists of the usual logical operators and brackets, the predicates $P$ and $\epsilon$, and the variables $v^m_1 (i \in \omega, m \leq n)$. The type $t(v^m_1)$ of a variable $v^m_1$ is $m$. Having defined $L_0$, we can now define the languages $L_n$ for $n > 0$ by the following recursion:

**DEF 3.3: FORMATION RULES FOR $L_n$ $(n > 0)$**

1. If $A$ is an atomic formula of $L_0$, then $A$ is a formula of $L_n$.
2. If $x^i$ and $y^{i+1}$ are variables of types $i$ and $i + 1$, respectively, where $i < n$, then
$x^i \in y^{i+1}$ is a formula of $L_n$.

(3) If $A$ and $B$ are formulas of $L_n$, then $\neg A$ and $(A \rightarrow B)$ are formulas of $L_n$.

(4) If $A$ is a formula of $L_n$ and $x^i$ is a variable of type $i \leq n$, then $\exists x^i A$ is a formula of $L_n$.

If we take languages to be sets of formulas, we can identify the hierarchy itself with the language $L = \bigcup_{n \in \omega} L_n$. The level $\ell(A)$ of a formula $A$ of $L$ is the smallest $m$ such that $A \in L_m$.

Several remarks about DEF 3.3 are in order. Clause (1) is stated in a general way to show that the construction of the hierarchy does not depend essentially upon the particular choice of $L_0$. Clause (2) makes the hierarchy strongly restrictive. The hierarchy is nested, since $L_n \subseteq L_m$ if $m > n$.

In addition to being the language of a Tarski hierarchy, $L$ is as well the language of simple type theory, with the extra predicate $P$. We will interpret each $L_n$ $(n > 0)$ over the unary predicative formulas of all $L_m$ $(m < n)$ so that $L$ becomes, not only the language of type theory, but also a model of CTT.

To effect this interpretation, we need a domain $D_n$ for each variable type $n$. We assume $D_0$ given and disjoint from $L$ and set $D_{n+1} = \{A \mid A$ is a predicative unary formula of $L$ and $\ell(A) = n\}$. Clearly each predicative unary formula
of $L$ belongs to one and only one of the domains $D_{n+1}$. Hence all the $D_n$ are disjoint. The hierarchy is therefore noncumulative, as well as predicative. Variable assignments for on the $D_n$ assign each variable of type $n$ a member of $D_n$. Truth relative to a variable assignment is defined by simultaneous recursion on the level and length of formulas:

**DEF 3.4:** INTERPRETATION OF $L$

For all formulas $A$ of $L$,

1. If $A$ is $P^0_x$, then $\models_I A[a]$ for all $a \in D_0$,
2. If $A$ is $\neg B$, then $\models_I A[a] \iff \models_I B[a]$,
3. If $A$ is $(B \rightarrow C)$, then $\models_I A[a] \iff \models_I B[a]$ or $\models_I C[a]$,
4. If $A$ is $\exists x^1 B$, then $\models_I A[a] \iff \models_I B[a,d]$ for some $d \in D_1$,
5. If $A$ is $\forall x^i y^{i+1}$, then for all $a \in D_i$, $b \in D_{i+1}$, $\models_I A[a,b] \iff \models_I b[a]$.

Clauses (2)-(4) ensure that truth is defined for every formula of level $n$ if it is defined for atomic formulas of level $n$. Now suppose all formulas of level $n$ have been interpreted. We use clause (5) to determine the truth conditions of atomic formulas of level $n+1$, i.e., formulas of the form $x^n \in y^{n+1}$. Now on any variable assignment which assigns $a$ to $x^n$ and $b$ to $y^{n+1}$, $a \in D_n$ and $b \in D_{n+1}$. Predicativity ensures that the free variable in $b$ is of type $n$, the
appropriate type for assignment to \(a\). On such an assignment, \(x^n \in y^{n+1}\) is true just in case \(a\) satisfies \(b\). Now by our supposition, the satisfaction conditions of \(b\) are already defined, so the definition is not circular like our previous attempt. And clause (5), accordingly, gives \(\in\) the meaning 'satisfies' just as intended. Our next aim is to show that this interpretation makes CTT true.

Since variables are now typed, identity can no longer be defined as in DEF 3.1. We need, instead, a definitional schema:

\[
\text{DEF 3.5: } x^n = y^n \text{ for } \forall z^{n+1} (x^n \in z^{n+1} \leftrightarrow y^n \in z^{n+1}).
\]

In the theory of types, this schema is regarded as a version of Leibniz' Law. The predicate \(=\) is taken to express absolute identity, and the variable \(z^{n+1}\) is thought to range over properties, attributes, or sets. For us, however, \(=\) expresses indiscernibility with respect to the predicative unary formulas of \(L_n\). (This amounts to indiscernibility in \(L_n\), as can be verified by a simple induction.)

The correlative notion of extensional equivalence is also defined schematically:

\[
\text{DEF 3.6: } x^{n+1} = y^{n+1} \text{ for } \forall z^n (z^n \in x^{n+1} \leftrightarrow z^n \in y^{n+1}).
\]

And the axiom of extensionality correspondingly becomes a schema:

\[
\text{EXT(S): } \forall x^{n+1} \forall y^{n+1} (x^{n+1} = y^{n+1} \leftrightarrow x^{n+1} \equiv y^{n+1}).
\]
Notice that extensional equivalence cannot be asserted of members of $D_0$ and that $\text{EXT}(S)$ gives identity conditions only for members of $D_n$ where $n > 0$. This corresponds to the metaphysical assumption of type theory that the objects of lowest type (the individuals) are ontologically different from the abstract objects of higher types and thus have different identity conditions. For us, it is simply a reflection of the fact that $L_0$ and the $L_n$ ($n > 0$) are different sorts of languages with different identity predicates.

The constructive theory of types is axiomatized by two schemata, $\text{EXT}(S)$ and the schema of comprehension:

\[
\text{COMP: } \exists y^{n+1} \forall x^n (x^n \in y^{n+1} \leftrightarrow A), \text{ where (1) } x^n \text{ is free in } A, (2) y^{n+1} \text{ is not free in } A, \text{ and (3) } A \text{ contains no variables of type greater than } n + 1 \text{ and no bound variables of type greater than } n.
\]

On our interpretation, COMP asserts the existence (possibility) of a formula of $D_{n+1}$ extensionally equivalent to $A$, for any instantiation of $A$'s free variables other than $x^n$. The following theorems show that both axiom schemata are true.

**THM 3.1:** If $A$ is an instance of $\text{EXT}(S)$, then $\models_1 A$.

**Proof:** Pick $a, b \in D_{n+1}$ and suppose $\models_1 x^{n+1} \equiv y^{n+1}[a, b]$.●
Then for some $c \in D_n$, $\models a[c]$ but $\not\models b[c]$, or vice versa. Assume the former. Where $x^n$ is the variable free in $b$, let $A$ be $\exists x^n(x^n \in y^{n+1} \land \sim b)$. Clearly $\models A[a]$ but $\not\models A[b]$. Thus $\models x^{n+1} \in z^{n+2}[a, A]$ but $\not\models y^{n+1} \in z^{n+2}[b, A]$. So $\models x^{n+1} = y^{n+1}[a, b]$. In the second case, the reasoning is exactly similar. Thus for all $a, b \in D_{n+1}$, $\models x^{n+1} = y^{n+1} + x^{n+1} \equiv y^{n+1}[a, b]$.

Conversely, pick $a, b \in D_{n+1}$ and suppose $\not\models x^{n+1} = y^{n+1}[a, b]$. Then there is some formula $A \in D_{n+2}$ such that $\models A[a]$ but $\not\models A[b]$. Since $A \in D_{n+2}$, $A$'s free variable $z^{n+1}$ occurs only to the right of $\in$ in $A$. Let $B$ and $C$ be the results, respectively, of replacing each subformula of the form $v^n \in z^{n+1}$ in $A$ by $a(v^n)$ and $b(v^n)$, where $a(v^n)$ is the result of replacing $a$'s free variable by $v^n$, and similarly for $b(v^n)$. (Some relettering of variables may be necessary to prevent binding of $v^n$ in $a$ and $b$ if $v^n$ already occurs in either.) Clearly, for all $c \in D_n$, $\models v^n \in y^{n+1}[c, a] \Rightarrow \models a[c]$, and similarly for $b$. So by induction on the construction of $A$ it follows that $\models B \Rightarrow \models A[a]$ and $\models C \Rightarrow \models A[b]$. But then since $\models A[a]$ and $\not\models A[b]$, we know that $\not\models B \leftrightarrow C$. But if $\models x^{n+1} = y^{n+1}[a, b]$, then $\models a(v^n) \leftrightarrow b(v^n)[c]$ for all $c \in D_n$, and so by substitutivity of equivalents, $\models B \leftrightarrow C$, and we have a contradiction. (Substitutivity of equivalents holds for formulas of $L$--provided their free variables are of the same type--by the usual sort of induction on formula construction.)
\[
\models^I x^{n+1} \equiv y^{n+1}[a,b]. \text{ Therefore, } \models^I x^{n+1} \equiv y^{n+1} + x^{n+1} = y^{n+1}[a,b], \text{ and this completes the proof of the theorem.}
\]

To show that the comprehension schema holds, we need the following definitions and lemmas.

**DEF 3.7:** If \( x^0 \) is a variable of type 0 and \( A \) is a formula, then the \( x^0 \)-elimination of \( A \) is \( \exists x^0 A \).

**LEMMA 3.2:** If \( B(z) \) is the \( x^0 \)-elimination of \( A(z,x^0) \), then for all \( c \in D_t(z) \) and \( a \in D_0 \),
\[
\models^I A(z,x^0)[c,a] \models^I B(z)[c].
\]

**Proof:** By **DEF 3.4** and induction on the construction of formulas, all pairs of members of \( D_0 \) are indiscernible in \( L_0 \). Hence for all \( a \in D_0 \) and \( c \in D_1 \), \( \models^I x^0 \in y^1[a,c] = \models^I x^0 \in y^1[b,c] \) for some \( b \in D_0 \), and \( \models^I P^0[a] = \models^I P^0[b] \) for some \( b \in D_0 \). It follows by induction on the construction of formulas that for any formula \( A(z,x^0) \), \( c \in D_t(z) \) and \( a \in D_0 \), \( \models^I A(z,x^0)[c,a] = \models^I A(z,x^0)[c,b] \) for some \( b \in D_0 \). The lemma is an immediate consequence.

**DEF 3.8:** Let \( A \) be any formula and \( B \) any unary formula.

We construct the \( x \)-elimination of \( A \) with respect to \( B \) as follows:

1. replace each subformula of the form \( x \in y \) in \( A \) by \( \exists x(\forall w(w \in x \leftrightarrow B) \land x \in y) \),

where \( w \) is \( B \)'s free variable; and
(2) replace each subformula of the form \( y \in x \) in \( A \) by \( B(y) \), where \( B(y) \) is the result of replacing \( B \)'s free variable by \( y \) (some relabeling perhaps being necessary to prevent binding of \( y \) in \( B(y) \)).

Step (1) eliminates all free occurrences of \( x \) before \( \epsilon \), and step (2) eliminates all free occurrences after \( \epsilon \).

**LEMMA 3.3:** Let \( x \) be a variable of type \( n + 1 \) and \( B \in D_{n+1}^1 \).

Let \( C(y) \) be the \( x \)-elimination of \( A(y, x) \) with respect to \( B \). Then for all \( a \in D_t(y) \),

\[
\models_I C(y)[a] \iff \models_I A(y, x)[a, B].
\]

**Proof:** (1) For all \( a \in D_t(y) \), \( \models_I x \in y[B, a] \iff (by \text{THM 3.1}) \models_I x \in z[b, B] \) for some \( b \in D_t(x) \) (where \( t(z) = t(x) \)) and \( \models_I x \in y[b, a] \iff \models_I \exists x(\forall w(w \in x \leftrightarrow B) \land x \in y)[a] \), where \( w \) is \( B \)'s free variable.

(2) For all \( a \in D_t(y) \), \( \models_I y \in x[B, a] \iff (by \text{DEF 3.4}) \models_I B(y)[a] \).

The lemma follows from equivalences (1) and (2) by induction on the construction of \( A(y, x) \).

**THM 3.4:** If \( B(z) \) is an instance of \( \text{COMP} \), then \( \models_I B(z)[a] \)

for all \( a \in D_t(z) \).

**Proof:** Let \( B(z) \) be an instance of \( \text{COMP} \). Then \( B(z) = \exists y^{n+1} \forall x^n(x^n \in y^{n+1} \leftrightarrow A(z, x^n)) \), where \( x^n \) is free in \( A \), \( y^{n+1} \)
is not free in $A$, and $A$ contains no variables of type higher than $n + 1$ and no bound variables of type higher than $n$.

Pick $a \in D_t(z)$. Let $c$ be the result of omitting all elements of $D_0$ from $a$, and let $C(w, x^n)$ be the result of taking the $x$-elimination of all free type 0 variables in $A$, except for $x^n$ (if $n = 0$). By repeated application of LEMMA 3.2, we can see that for all $b \in D_n$, $\models C(w, x^n)[c, b] = A(z, x^n)[a, b]$. Now let $D(x^n)$ be the result of taking the $w_1$-elimination of each variable $w_1$ among the $w$ with respect to the corresponding object $c_1$ among the $c$. Then by repeated application of LEMMA 3.3, we know that for all $b \in D_n$, $\models D(x^n)[b] = C(w, x^n)[c, b] = A(z, x^n)[a, b]$.

Now $D(x^n)$ is a predicative unary formula and so $D(x^n) \in D_{n+1}$. Therefore, $\models \forall x^n(x^n \in y^{n+1} \leftrightarrow A(z, x^n)) [D(x^n), a]$ and consequently $\models B(z)[a]$.

As a set theory, CTT lacks the theorem-proving power needed for most applications. To be adequate even for simple arithmetic, for example, it requires supplementation by the axiom of infinity, which on our interpretation asserts that $D_0$ has infinitely many discernible members. This axiom is false unless we make important alterations in our hierarchy and its interpretation—including the addition of predicates of $L_0$ so that the members of $D_0$ become discernible, and the stipulation that $D_0$ is infinite. Even with the axiom of infinity, however, CTT is inadequate for higher mathematics, since it permits no proof of the least
upper bound theorem. This theorem is also false on our interpretation of $L$.

Any hierarchy which models a powerful set theory like ZF must evidently by more sophisticated than $L$. The first step toward defining a hierarchy for a system more powerful than CTT was taken by C. S. Chihara in his book *Ontology and the Vicious-Circle Principle*. Chihara showed in effect that Wang's set theory $\Sigma_\omega$ has a model consisting of unary formulas hierachically arranged, though he did not show that these formulas can be construed as elements of a Tarski hierarchy. Wang's system apparently permits proof of most important mathematical results. Chihara, Lin, and Schaffter report that in a revised version of the system "analogues of the classical theorems have been established up through the basic results of Lebesgue integration, including the proof that the space of square integrable functions is a complete (...in an appropriate sense), separable, inner produce space, i.e., a Hilbert space." Chihara's interpretation of $\Sigma_\omega$ employs modal quantification of a more limited sort than that discussed in Chapter 2. His reading of the quantifiers is explained in the following passage:

...given a sentence $(\exists a)\phi$, one can specify an inductive rule for generating all the sentences of the form $\phi a/\beta$, where $\beta$ is a constant of the same order as $a$. $(\exists a)\phi$ will then mean: If one were to follow this rule of construction without end, one would eventually construct a true
sentence. One could also interpret the
sentence to read: It is possible to con­
struct a true sentence by following this
rule of construction (CHIHARA [1973],
p. 191).

Chihara's "rules of construction"--together with his rules
of interpretation (see pp. 216-17)--are analogous to what we
called "grammatical rules" in Chapter 2. His interpretation
of $\Sigma_w$, however, is somewhat nonstandard. Though he reads
his quantifiers modally, the interpretation is in effect
substitutional. Moreover, Chihara's notion of what is
admissible as a rule is narrower than ours. The formulas
which he counts as constructible are all generable from a
finite stock of primitives by recursive rules in a finite
number of steps. Furthermore, his interpretations are given
by rules which make no reference to "totalities" other than
those constructively specifiable by further rules. And
finally, Chihara requires that

\[
\text{[e]ach open sentence constructed must be well-defined over its argument range in the following sense: If } \phi \text{ is a member of the argument range of } \psi, \text{ then } \psi \text{ is either true of } \phi \text{ or false of } \phi \text{ (but not both) (p. 186).}
\]

We shall not insist upon any of these requirements.

Our more liberal understanding of the notion of a
grammatical rule is thus a considerable departure from
Chihara's constructivistic attitude. Accordingly, it
would be misleading for us to read quantifiers as Chihara
does in terms of modal or counterfactual talk about constructibility. We are concerned, not simply with constructively generable and interpretable formulas, but with possible formulas of any sort—constructible or not. For that reason, we prefer the reading "there is possible..." or "there could be..." for 3. This increased latitude permits extension of the analysis of sets as open formulas to systems as powerful as ZF. It also attenuates the notion of an open formula.

There is an obvious trade-off: Chihara takes a conservative attitude about what counts as a constructible formula and winds up with a relatively weak set theory; we allow quantification over a wider class of possible tokens and extend the analysis to more powerful theories.

It may be objected that the objects permitted by these liberalizations are not formulas in the strict sense, so that our analysis is not really a reduction of sets to formulas. There is no need to quibble about the use of a term. If the objects we shall soon quantify over are not all formulas in the strict sense, at least they have definite and easily understandable similarities to objects which everyone agrees are formulas. Moreover, many of them are formulas in the strict sense—indeed, actual formulas.

There are several reasons for our liberality. First, it allows a natural account of actual set-theoretic practice. When mathematicians and others use set theory, they use ZF
or a closely related system, not $\Sigma^*_\omega$. Our purpose here is to account for what they are doing, rather than to say what they should be doing. We make the methodological assumption that what they are doing is intelligible as is, and then try to fit our analysis to it. Ours is not a reform program.

Furthermore, though the tokens we permit are sometimes more exotic in construction or interpretation than Chihara's, still many of them are actual or could be made actual by finite means. It is this link with actuality, with what the set theorist actually says and writes, not constructivity, which we regard as essential to the solution of the metaphysical and epistemological problems of set theory. Since our purpose is to shed light on these problems, it is therefore sufficiently served by our more lenient notion of possibility.

Finally, our stretching of constructibility into possibility is justified by the desire to see just how much stretching in this direction is necessary to accommodate classical set theories. Even if it does not ultimately lead to an enlightening analysis, the pursuit of this question may be of some intrinsic interest. Of course, we claim that it is of much more interest than this, and that it does provide an enlightening analysis; but justification of these claims must await the detailed working-out of the analysis itself.
A second step toward identification of set theoretic universes with Tarski hierarchies was taken in McDermott (1977). Unlike Chihara, McDermott locates the open formulas of the domain of his set theory ITT within a Tarski hierarchy. But he avoids the weaknesses of the simple Tarski hierarchy of CTT by some ingenious modifications. For one thing, the hierarchy is predicate-typed. Each language $L_n$ ($n \in \omega$) of the hierarchy contains the predicates $\epsilon_m$ ($m \leq n$), whose intuitive interpretation is "satisfies-in-$L_m$." Dispensing with typed variables permits interpretation of all the languages over a single domain containing the unary of each $L_n$, plus some additional formulas, about which more will be said momentarily. $L_0$ is interpreted so that no formula satisfies-in-$L_0$ any other (i.e., $x \epsilon_0 y$ is false of all pairs in the domain). But then $\sim x \epsilon_0 x$, for example, is true of all objects in the domain. In interpreting $L_1$, therefore, McDermott stipulates, among other things, that each formula in the domain satisfies-in-$L_1$ the formula $\sim x \epsilon_0 x$. The interpretation process continues in this manner, more and more formulas being satisfied-in-$L_n$ by others as $n$ increases. Since each $L_n$ quantifies over the entire domain, however, even at the lowest level there is quantification over an infinity of objects. McDermott adds at the lowest level predicates sufficient to make all these formulas discernible. The result is the truth of the axiom of infinity.
The second innovation is that after interpreting $\varepsilon_n$ and before moving to the interpretation of $\varepsilon_{n+1}$, McDermott adds to $L_n$ and interprets a host of additional predicates—namely, all those inductively definable from $\varepsilon_n$. The resulting complexity permits strengthening of the comprehension principle in such a way that the least upper bound theorem becomes provable, and with it most of classical mathematics. 8

McDermott's predicate-typed hierarchy with its added inductive predicates is the language of a theory he calls SF(a). The set theory ITT which he actually proposes, though it quantifies over this hierarchy, employs a different language. The language of ITT is just the language of simple type theory (i.e., the language $L$ in which we formulated CTT, minus the predicate $P$). ITT's comprehension principle is identical to COMP, except that in place of Clause (?) is has simply "A contains no variables of type greater than $n + 1$." This strengthened comprehension principle, EXT(S), and the axiom of infinity are both true on the intended interpretation.

ITT is a stronger theory than $\Sigma_\omega$, though it is weaker than classical set theories. In $\Sigma_\omega$ the set of all real numbers cannot be shown to exist and it is impossible to define functions over all the reals. This means that considerable effort is required to establish analogues of the theorems of higher mathematics in $\Sigma_\omega$, though Chihara and his
colleagues seem to have solved most of the difficulties. Nevertheless, many standard proofs require modification, and the development of mathematics in $\Sigma_\omega$ is correspondingly inelegant. Not so for ITT. The set of all reals is definable, and the development of mathematics proceeds roughly along the lines of its development in simple type theory.

Not surprisingly, then, ITT, unlike $\Sigma_\omega$, seems not to be constructive. There appears to be no way to state general rules for the interpretation of the inductive predicates added at each stage without reference to totalities which are not specifiable by a constructive rule. Thus the interpreted unary formulas of SF(a), while possible, appear not to be constructible, at least as McDermott intends to interpret them. Further steps away from constructibility will be taken in the next chapter.
After the discovery of Russell's antinomy in 1901, set theory branched off in two directions. Russell himself sought the solution in an elaborate ramified theory of types, of which the system CTT of the last chapter is a weaker and simpler descendent. Theories of types avoid the antinomies by fragmenting both the language and its domain of quantification into levels or types. For this reason they are inelegant and unwieldy in use.

Zermelo and his colleagues, working independently at Gottingen, hit upon a simpler and more satisfactory solution. Stratification of the universe alone, they discovered, was apparently sufficient to purge the antinomies. If sets themselves were conceived as "constructed" in stages or layers, then stratification of the language was unnecessary. An untyped first-order language, elegant and simple to use, would do. It was Zermelo's idea, naturally enough, which eventually gained wide acceptance and now flourishes as Zermelo-Fraenkel set theory. Theories of types have attracted comparatively little interest.
Yet if our analysis is correct, these two ways of doing set theory have much in common; in particular, they are both ways of talking about language. This is not too difficult to see in Russell's case. The language $PM$ of *Principia Mathematica* is closely related to the language $L$ of Chapter 3. $PM$ has typed variables, but instead of employing a primitive satisfaction predicate it simply allows higher level variables to occupy the predicate position with respect to lower level ones. Thus where we would write $x^1 \epsilon y^{i+1}$ in $L$, Russell has $\phi!x$, where $\phi$ and $x$ are both variables. (Russell avoids mention of the order of a variable by a convention of systematic ambiguity; $x$ and $\phi$ are variables of unspecified order, and the exclamation mark indicates that the order of $\phi$ is exactly one higher than that of $x$.) What we regard in $L$ as the primitive satisfaction predicate is a defined symbol of $PM$.

The variables of $PM$ other than those of the first order were to range over what Russell called "propositional functions." As Quine and others have noted, Russell's pronouncements about propositional functions are equivocal. He alternatively thinks of them as open formulas and as properties or attributes. If we adopt the former reading and, in addition, regard the satisfaction predicate as primitive instead of allowing predicate variables, then the relation of $PM$ to a Tarski hierarchy becomes clear. It is, however, a more complex hierarchy than those discussed in
Chapter 3. Russell allows quantification over n-adic propositional functions, the equivalent of which in $L$ would be to allow quantification over open formulas in any number of free variables. To accommodate the new formulas in $L$, we would have to add an infinite number of new satisfaction predicates (one for each number of free variables) or, alternatively, make $\epsilon$ a variably polyadic predicate like the predicate $T$ mentioned in Chapter 1.

Moreover, considered as a Tarski hierarchy, $PM$ quantifies over impredicative formulas as well as predicative ones. This feature can be incorporated into a Tarski hierarchy, but only at the cost of "ramification"—the division of variables into both orders and types—and a tremendous increase in complexity.

Finally, however, Russell added an axiom to the theory of $PM$ which is false if $PM$ is interpreted as a Tarski hierarchy in any straightforward way. This is the axiom of reducibility, which stipulates that any propositional function taking a given object as value has a coextensive or equivalent function of order one higher than that of the object. The axiom in effect allows objects of higher order to be considered as being of lower order and thus disturbs the strict sequence of construction which characterizes a Tarski hierarchy. Without the axiom, however, both Russell's universe and his language can be regarded as Tarski hierarchies—indeed, as the same hierarchy.
Like Russell's universe, the universe of ZF shares important features with Tarski hierarchies. Like a Tarski hierarchy, it is conceived as being "constructed" in an ordered sequence of stages, and its members can be regarded as belonging to various levels. The sets of each of these levels have only sets of lower levels as elements, and this corresponds to the fact that in a Tarski hierarchy formulas or one level are usually interpreted over and always satisfied only by formulas of lower levels. However, Zermelo's universe is a much loftier structure than either Russell's type hierarchy or $L_S$. The constructive process does not approach an ultimate limit at $\omega$, but continues through every ordinal level.

Zermelo's language, too--or, more precisely, the language of ZF--differs in important respects from Russell's. Unlike $PM$ or $L$, it is not identifiable with the hierarchy, but is instead a single separate language which is to be interpreted over the entire hierarchy. In fact, it is just the language $L_S$ of Chapter 3.

What we shall show in this chapter is that $L_S$ can be interpreted over the possible unary formulas of a hierarchy of languages not unlike a Tarski hierarchy in such a way that all the axioms of ZF are true. That is, we shall show that the possible formulas of such a hierarchy form a model for ZF. This proof cannot be carried out without strong assumptions, since the construction of a model of ZF proves
ZF's consistency; and, as is well-known, the consistency of ZF cannot be proved even in ZF if ZF is consistent.

The strong assumption we need is just that ZF is consistent. Given this assumption, we can then choose an arbitrary model of ZF. Within this model we employ a technique akin to Gödel-numbering, letting certain objects of the model represent the vocabulary and possible expressions of both the set-theoretic object language $L^S$ and the hierarchy $L^{ZF}$ over which it quantifies. We then define $L^{ZF}$ by stating formation and interpretation rules (DEF 4.3 and 4.6) which pick out the subclass of the representatives of possible expressions of $L^{ZF}$ which is to count as the class of representatives of formulas and define the satisfaction relation over this class.

The structure of the class thus defined is closely akin to the structure of another subclass of our arbitrary model—Gödel's constructible hierarchy $L$. This second class is known to model both ZF and Gödel's axiom of constructibility $V = L$. To indicate this similarity of structure, we define a "proxy function" $\eta$, which maps members of the first class onto members of the second. Using $\eta$, we can then show that when $L^S$ is interpreted over each class, every formula true on the one interpretation is also true on the other (COROLLARY 4.10). It follows that our class of representatives of formulas models $ZF + V = L$. Since $V = L$ implies the axiom of choice and the generalized continuum
hypothesis, our hierarchy also models these axioms.

In Chapter 2 we noted that possibility for formula tokens is relative to a set of grammatical conventions. Our statement (in DEF 4.3 and 4.6) of formation and interpretation rules for a hierarchy modelling ZF will therefore serve to justify our claim that tokens belonging to a hierarchy which models ZF are possible. This justification, of course, ultimately rests upon the assumption of ZF's consistency, but for that assumption mathematical practice of the past seventy years provides convincing inductive evidence.

For purposes of the proof, we shall take ZF as our metatheory. There is no circularity in this procedure, as we noted at the end of Chapter 2. In assuming that ZF has a model and operating set-theoretically within the model, we make no commitment either to the sorts of objects which the model contains or to the metaphysical nature of sets or classes of them. Our proof will show that if ZF is consistent, then if has a model which is a hierarchy of languages. Once this is understood, it becomes obvious that the original model could have been a hierarchy of possible tokens, as well as a universe of platonic sets. Moreover, it becomes obvious that the set-theoretic quantifiers of our metalanguage could also range over possible tokens rather than over sets. Initially, however, we simply regard our metatheory as unanalyzed in the manner explained in Chapter 2.
We saw earlier that Tarski hierarchies can be constructed in a great variety of ways. Now we must consider what is needed to construct a hierarchy which models ZF. Although, because of a slight alteration in the interpretation of the predicate $\in$, $L^{ZF}$ will not, strictly speaking, be a Tarski hierarchy, it will be easiest to see the connection between $L^{ZF}$ and genuine Tarski hierarchies if we begin by trying to construct a Tarski hierarchy which models ZF. In the course of this construction, the reason for the alteration just mentioned will become evident.

We already know that our hierarchy must be transfinite. And, since sets of one level in the universe of ZF can be members of sets of all higher levels, its formulas should be satisfiable by formulas of all lower languages, not just the next lowest. The hierarchy will therefore be cumulative. There is no apparent advantage either to predicate-typing or to variable-typing, and either would do for our purposes. We arbitrarily choose the latter.

It is not immediately obvious, either, whether the languages of the hierarchy should be nested or discreet. In practice, nested variable-typed hierarchies are inconvenient, since their member languages have variables of many types. In a discreet variable-typed hierarchy, on the other hand, each language has a single characteristic variable type. A discreet hierarchy will prove adequate for our purposes.
Finally, we must decide on the details of the construction of the languages themselves. It will prove essential that each language contain constants sufficient to name each of the unary formulas of all languages of lower levels. Because the hierarchy is transfinite, this stock of constants will enrich the variety of unary formulas available at infinite levels. The number of constants available at very high levels will eventually become nondenumerable and thereby increase the expressive power of these languages in such a way that the hierarchy will come to mirror Gödel's constructible universe $L$ of sets. It is this mirroring which enables us to show that the hierarchy models ZF.

The necessary constants can be constructed through the use of a quotation operator, which we shall represent by corner quotes '($I$ and ')'. Since the hierarchy is transfinite, nesting of formulas containing constants inside new constants eventually results in constants of infinite length, i.e., containing infinitely many symbols. This would be unacceptable if we were trying to reduce sets to actual or even physically possible tokens, but there is nothing in the concept of possibility which we are employing that prevents the possibility of infinite tokens. The role of the constants will become clearer as we proceed, and at the end of this chapter we shall consider a way of interpreting the hierarchy without them.
We now assume that ZF has a model and begin construction of the hierarchy within it. The hierarchy $L^{ZF}$ will be the union of languages $L^{ZF}_a$ for each ordinal $\alpha > 0$. (Beginning with 1 rather than 0 is a technicality which will prove useful later.) The vocabulary for each $L^{ZF}_a$ consists of the usual logical particles and brackets, the predicate $\epsilon$, quotation marks ' and ', and the variables $v^\beta_1$ for $0 < \beta \leq \alpha$ and $i \in \omega$. Only the variables $v^\alpha_1$ actually function as variables in $L^{ZF}_a$. The others are used only for the construction of constants. When referring to variables we shall, where convenient, omit reference to subscripts, superscripts, or both. Where notation is used without definition, it will officially be that of DRAKE (1974), but in all cases such notation is fairly standard.

To represent elements of the vocabulary of each $L^{ZF}_a$, we arbitrarily choose certain sets of our assumed model.

DEF 4.1: $\epsilon$ for 1 (for 5
~ for 2 ) for 6
$\to$ for 3 (a) for <7,a>, for any object a
$\exists$ for 4 $v^\alpha_1$ for $\omega_\alpha + 1$, for $i \in \omega$ and $\alpha > 0$

Concatenation is represented by juxtaposition in a sequence. Thus when we write $\exists v^1_0 \sim v^1_0 \epsilon v^1_0'$, for example, we are referring to the sequence <4, $\omega_1$, 2, $\omega_1$, 1, $\omega_1$>. In general, if $a_1$, ..., $a_n$ are elements of our vocabulary, we put
DEF 4.2: \( a_1 \ldots a_n \) for \( \langle a_1, \ldots, a_n \rangle \).

As in Chapter 3, we define the type \( t(v) \) of a variable \( v \) as the ordinal designated by its superscript. In view of DEF 4.1, the type of a variable is its cardinality. And, as before, we interpret the base language \( L^ZF_1 \) over an arbitrary nonempty domain \( D_1 \). It will be convenient—though, as we shall see, not necessary—to have a constant in each \( L^ZF_\alpha \) for each object in \( D_1 \). This can be done by allowing the quotation operator to apply to the objects of \( D_1 \) to form their names. This is merely a technical expedient which will prove more efficient than the introduction of primitive names. We assume that \( D_1 \) is disjoint from the set of symbols and sequences of symbols characterized by DEF 4.1 and 4.2. Supposing \( D_1 \) given, we can define the languages \( L^ZF_\alpha \) by induction on \( \alpha \).

DEF 4.3: FORMATION RULES FOR \( L^ZF_\alpha \) (\( \alpha > 0 \))

(1) An object \( a \) is a constant of \( L^ZF_\alpha \) if
\( a = \langle b \rangle \), where \( b \in D_1 \) or \( b \) is a formula of \( L^ZF_\beta \) for some \( \beta \) (\( 0 < \beta < \alpha \)).

(2) An object \( a \) is a term of \( L^ZF_\alpha \) if \( a \) is a constant of \( L^ZF_\alpha \) or \( a \) is a variable of type \( \alpha \).

(3) \( L^ZF_\alpha \) is the smallest class \( S \) such that

(1) if \( s, t \) are terms of \( L^ZF_\alpha \), then
\( s \in t \in S \),
(ii) if $A, B \in S$, then $\sim A, (A \rightarrow B) \in S$,

(iii) if $A \in S$ and $x$ is a variable of type $a$, then $\exists x A \in S$.

(4) $A$ is a formula of $L^Z_F$ if $A \in L^Z_a$.

The terms atomic formula, free variable, unary formula, and so on are defined as in Chapter 1. No variable of type less than $a$ is considered to be a variable of $L^Z_a$, so that no variable within a constant of $L^Z_a$ counts as a variable of $L^Z_a$. Since constants are themselves sequences and not sub-sequences of formulas, even though constants may be of infinite length, all formulas are of finite length.

('Length' here simply refers to the length of the sequence representing an expression.) We also define

DEF 4.4: $L^Z_F = \bigcup_{a > 0} L^Z_a$.

Like $L$, $L^Z_a$ is to be interpreted over a hierarchy of domains $D_a$. For each $L^Z_a$, $D_a$ consists of the members of $D_1$ together with unary formulas of all lower languages. We assume $D_1$ given and define the domains $D_a$ as follows:

DEF 4.5: (1) For all $a > 1$, $D_a = D_1 \cup \{a \mid a$ is a unary formula of $L^Z_F$ for some $\beta < a\}$.

(2) $D = \bigcup_{a > 0} D_a$.

Clearly if $a > \beta > 0$, $D_\beta \subset D_a$. 

Now we could make $L^{ZF}$ into a Tarski hierarchy by interpreting $\in$ in each succeeding language as the satisfaction predicate for the languages below it. However, such a hierarchy would not model ZF in any straightforward way. This is because throughout such a hierarchy extensionally equivalent formulas—the ones which must be indiscernible if they are to "appear" as sets—are discernible by higher languages. Consider, for example, the formulas $x^1 \in x^1$ and $x^2 \in x^2$. Since no member of a Tarski hierarchy satisfies itself, neither would be satisfied by any object in the hierarchy; so they would be extensionally equivalent. But the formula $\neg \exists x^3 (y^3 \in x^3)$ would discern them, since, though there would be no object in $D_3$ satisfied by $x^2 \in x^2$ (since $D_3$ contains no languages higher than $L^{ZF}_2$), there would be objects (e.g., $\neg x^2 \in x^2$) satisfied by $x^1 \in x^1$. The result is that, though $x^1 \in x^1$ and $x^2 \in x^2$ are extensionally equivalent and hence ought to be counted as the same set, the first is a "member" of $\neg \exists x^3 (y^3 \in x^3)$, while the second is not. This problem did not arise in $L$ because $L$ was noncumulative, but it is unavoidable in $L^{ZF}$, since cumulativity is required for a model of ZF.

The problem can be solved, however, by a simple modification of the interpretation of $\in$. Instead of reading $x \in y$ as 'x satisfies y', we could interpret it as 'something extensionally equivalent to x satisfies y'. On this reading $x^1 \in x^1$ and $x^2 \in x^2$ are again not satisfied by any object, but
since they are extensionally equivalent $\neg \exists x^3(y^3 \in x^3)$ is true of both. This modification, as we shall see, restores extensionality; all extensionally equivalent objects in the domain of each $L_a^{ZF}$ become indiscernible by $L_a^{ZF}$ and hence "appear" as sets. Though each language of the hierarchy thereby ceases to be a theory of satisfaction for those below it, it nevertheless is something closely related; we might call it a theory of "satisfaction by ancestral equivalents."

We now define truth relative to variable assignment for each $L_a^{ZF}$ according to this modified interpretation of $\epsilon$. The definition is by simultaneous induction on $\alpha$ and on the construction of formulas.

**Def 4.6: Interpretation of $L_a^{ZF}$**

1. For all $b \in D, a \in D_1, \models_{ZF} b[a]$.
2. For any formula $A$ of $L_a^{ZF}$ without constants and with $n \in \omega$ free variables, and for any $a = a_1, \ldots, a_m (m \geq n)$ in $D_a$,
   
   (i) if $A$ is $x^a \in y^a$, then $\models_{ZF} A[a] \iff$
   
   there exists $b \in D_a$ such that for all $c \in D_a, \models_{ZF} b[c] \iff \models_{ZF} a_1[c]$ and $\models_{ZF} a_2[b]$,
   
   (ii) if $A$ is $\neg B$, then $\models_{ZF} A[a] \iff \not\models_{ZF} B[a]$,
   
   (iii) if $A$ is $(B \rightarrow C)$, then $\models_{ZF} A[a] \iff$
(iv) if $A$ is $\exists B(x,y)$, then $\models_{ZF} A[a] = \models_{ZF} B(x,y)[b,a]$ for some $b \in D_a$.

(3) For any formula $A(c',x^a)$ of $L_{ZF}^a$ with constants $c'$ and possibly free variables $x^a$, and for any $a \in D_a$, $\models_{ZF} A(c',x^a)[a] = \models_{ZF} A(y^a,x^a)[c,a]$, where $A(y^a,x^a)$ is the result of replacing each $c_i'$ in $A(c',x^a)$ by $y_i^a$, $y^a$ being distinct variables not already in $A$.

(4) For any formula $A$ of $L_{ZF}^a$, if $a$ contains some $a_i$ not in $D_a$, then $\not\models_{ZF} A[a]$.

We shall call the interpretation of $L_{ZF}^a$ given by DEF 4.6 $L_{ZF}^a$'s standard interpretation. The definition itself requires some explanation. Clause (1) stipulates that no object in $D_1$ is satisfied by any assignment of members of $D$ to its free variables. This is appropriate, since the objects in $D_1$ have no free variables and are not satisfied by anything (they are arbitrary objects, distinct from the formulas of $L_{ZF}^a$). Defining satisfaction on an assignment for these objects is a technicality which makes subsequent definitions simpler.

If we have defined truth on an assignment for each of the objects in $D_a$, Clauses (2)-(4) tell us how to define truth on an assignment for the language $L_{ZF}^a$. Subclause (1)
stipulates that $x^a \in y^a$ is true with $a_1$ assigned to $x^a$ and $a_2$ assigned to $y^a$ for $a_1, a_2 \in D^a$, if something extensionally equivalent to $a_1$ satisfies $a_2$. This gives the interpretation sketched earlier. Now, with truth defined for atomic formulas without constants, Subclauses (ii)-(iv) and Clause (3) define truth on an assignment in $D^a$ for the rest of $L^F_\alpha$'s formulas.

Clause (4) stipulates that no sentence of $L^F_\alpha$ is true on any assignment which includes objects not in $D^a$. This is necessary to prevent expressions of the form $\vdash ZF A[a] \gamma$ from being undefined when members of $a$ belong to domains higher than that over which $A$'s variables range. Without Clause (4), expressions occurring in Subclause (1) would sometimes be undefined. Clause (4) is also presupposed by the following definitions:

**DEF 4.7:** For all $a \in D$, $a$ is **null** if for all $b \in D$,

$\not\models_{ZF} a[b]$.

**DEF 4.8:** For all $a, b \in D^a$, $a$ and $b$ are **extensionally equivalent** over $D^a$ (written $a \cong b^\gamma$) if for all $c \in D^a$, $\models_{ZF} a[c] \iff \models_{ZF} b[c]$.

**DEF 4.9:** For all $a, b \in D$, $a$ and $b$ are **extensionally equivalent** (written $a \sim b^\gamma$) if for all $c \in D$,

$\models_{ZF} a[c] \iff \models_{ZF} b[c]$.

Extensional equivalence and extensional equivalence over $D^a$.
are both equivalence relations. We can now verify that at each level \( a \), objects which are extensionally equivalent over \( D_a \) are indiscernible in \( L^ZF_a \). This indicates that \( L^ZF_a \) "sees" the members of \( D_a \) as sets, so that the problem of equivalent discernibles is avoided.

**THM 4.1:** For \( a > 1 \) and \( a, b \in D_a \), if \( a \equiv b \), then \( a \) and \( b \) are indiscernible in \( L^ZF_a \) on its standard interpretation.

**Proof:** Since \( \epsilon \) is \( L^ZF_a \)'s only predicate, it will suffice to show that if \( a \equiv b \), then for all \( c \in D_a \), \( \models_{ZF} x^a \epsilon y^a[c,a] \Rightarrow \models_{ZF} x^a \epsilon y^a[c,b] \) and \( \models_{ZF} x^a \epsilon y^a[a,c] \Rightarrow \models_{ZF} x^a \epsilon y^a[b,c] \).

Assume \( \models_{ZF} x^a \epsilon y^a[c,a] \). Then \( \models_{ZF} b[d] \) for some \( d \in D_a \) such that \( d \not\equiv c \). Since \( a \equiv b \), \( \models_{ZF} b[d] \). So, since \( d \not\equiv c \),

\( \models_{ZF} x^a \epsilon y^a[c,b] \). The converse is similar.

Suppose, on the other hand, that \( \models_{ZF} x^a \epsilon y^a[a,c] \).

Then \( \models_{ZF} c[d] \) for some \( d \in D_a \) such that \( d \not\equiv b \). But then \( \models_{ZF} x^a \epsilon y^a[b,c] \). Again, the converse is similar.

**COROLLARY 4.2:** For \( a > 1 \), \( a, b, c \in D_a \), and \( A \in L^ZF_a \), if \( a \equiv c \), then \( \models_{ZF} A[b,a] \Rightarrow \models_{ZF} A[b,c] \).

**Proof:** Immediate from THM 2.1 and THM 4.1.

To show that \( L^ZF \) is a model of ZF, we must interpret \( L^S \) over it in such a way that the axioms of ZF are true.

If we interpret the predicate \( \epsilon \) of \( L^S \) as "satisfies," we encounter the same problem we faced with \( L^ZF \)--certain
extensionally equivalent objects will be discernible. The formulas \( x^1 \in x^1 \) and \( x^5 \in x^5 \), for example, are both null and hence extensionally equivalent. But the first satisfies \( \sim x^2 \in x^2 \), while the second does not. Since the disease is similar, the same medicine is prescribed. We shall interpret the \( \epsilon \) of \( L^S \), like the \( \epsilon \) of each \( L^ZF \), as expressing satisfaction by an extensional equivalent. Since \( L^S \) is the language of \( ZF \), this is the meaning we are giving to the "membership" predicate of \( ZF \).

As with \( L^ZF \), we represent the formulas of \( L^S \) by sequences of sets. The symbols of \( L^S \) are the usual logical particles, brackets, \( \epsilon \), and the variables \( v_i \) for each \( i \in \omega \). We define the logical particles, brackets, and \( \epsilon \) as in DEF 4.1 and in addition let \( v_1 = 7 + i \). The formation rules for \( L^S \) are those for a classical first-order language. We define \( L^S \) as the set of formulas of \( L^S \) and interpret it as follows:

**DEF 4.10: INTERPRETATION OF \( L^S \)**

For any formula \( A \) of \( L^S \) in \( n \in \omega \) free variables, and for any \( a = a_1, \ldots, a_m \) (\( m \geq n \)) in \( D \),

1. if \( A \) is \( x \in y \), then \( \models S A[a] \iff b \sim a_1 \)
   and \( \models ZF a_2[b] \) for some \( b \in D \),
2. if \( A \) is \( \sim B \), then \( \models S A[a] \iff \models S \neg B[a] \),
3. if \( A \) is \( (B \rightarrow C) \), then \( \models S A[a] \iff \models S \neg B[a] \)
   or \( \models S C[a] \),
(4) if \( A \) is \( \exists x B(x,y) \), then \( \models_S A[a] \models_S B(x,y)[b,a] \) for some \( b \in D \).

DEF 4.10 gives the **standard interpretation** for \( L^S \).

We aim to show that the sets of ZF—or, more precisely, those of Gödel's constructible universe, which is a model of ZF—can be reduced to the possible unary formula tokens of \( L^ZF \). The members of \( D_1 \) play no essential role in this reduction. Since they are all null, \( L^S \) will "see" all of them indiscernibly as the null set. And since \( L^ZF \) contains null formulas, it already contains objects to which the null set can be reduced. The objects of \( D_1 \) are included in \( D \) because it is technically convenient to have a nonempty domain of objects over which to interpret \( L^ZF_1 \). We could begin with a language \( L^ZF_0 \), interpret it over the empty domain, and then construct the rest of the hierarchy as we have done. Each of the formulas of \( L^ZF_0 \) would be null, and so these formulas would do the work that \( D_1 \) does in our present hierarchy. But a null domain would require special truth conventions, so we have used \( D_1 \) instead.

We now prove a theorem analogous to THM 4.1 for \( L^S \).

**THM 4.3:** For all \( a, b \in D \), if \( a \sim b \), then \( a \) and \( b \) are indiscernible in \( L^S \) under its standard interpretation.

**Proof:** Similar to the proof of THM 4.1.
COROLLARY 4.4: For all $a, b, c \in D$ and $A \in L^S$, if $a \sim c$, then $\models_s A[b,a] \iff \models_s A[b,c]$.

Proof: Immediate from THM 2.1 and THM 4.3.

COROLLARY 4.5: $\models_s \forall x \forall y (x = y \iff x \equiv y)$

Proof: Immediate from THM 4.3 by DEF 3.1 and DEF 3.2.

COROLLARY 4.5 shows that the axiom of extensionality is true on the standard interpretation of $L^S$. The objects in $D$ are therefore behaving like sets from $L^S$'s point of view. Since all extensionally equivalent members of $D$ are indiscernible, this means that if we think of $L^S$ as a hierarchy of possible tokens, rather than as sequences of sets, these tokens, too, will be viewed by $L^S$ as sets, since equivalent ones will be discernible. It remains to prove that our interpretation makes the other axioms of ZF true.

Working still within our assumed model of ZF, we now define a submodel, Gödel's constructible universe $L$. We shall use the letter 'E' to represent the membership relation in our assumed model. That is, $E$ is the class $\{<a,b> \mid a \in b\}$. The definition of $L$ and consequent proof that $<L,E \models L$ models ZF and Gödel's axiom of constructibility $V = L$ are well-known and will not be discussed in detail here. (The proof first appeared in Gödel [1940], and an especially perspicuous version is given in Devlin [1973]. A thorough discussion of these matters also appears
in DRAKE [1974], Chapter 5. The reader is referred to these works for details.) After defining L, we shall show that any sentence of $L^S$ true on the standard interpretation is also true in the model $<L, E \models L>$, and vice versa. It will then follow that ZF and $V = L$ are true on the standard interpretation of $L^S$. We require the following definitions:

DEF 4.11: A set $a$ is semidefinable over a set $X$ if there are $c \in X$ and $A(y,x) \in L^S$ such that
$$a = \{b \in X \mid <X, E \models X> \models A(y,x)[c, b]\}.$$

DEF 4.12: Def($X$) is the set of all sets semidefinable over $X$.

We now define the constructible universe $L$ in stages $L_\alpha$, by induction on $\alpha$.

DEF 4.13: $L_0 = \emptyset$
$$L_{\alpha+1} = \text{Def}(L_\alpha)$$
If $\alpha$ is a limit ordinal, $L_\alpha = \bigcup_{\beta < \alpha} L_\beta$.
$$L = \bigcup_{\alpha \in \text{On}} L_\alpha$$

Note that if $\alpha > \beta$, $L_\beta \subseteq L_\alpha$.

We shall often wish to talk about the model structures $<L_\alpha, E \models L>$ and $<L, E \models L>$. Where the ambiguity generates no confusion, these will be referred to simply as $L_\alpha$ and $L$, respectively.
Our next task is to define the function $\delta : D \to L$. To each $a \in D$, $\delta$ assigns the set for which $a$ is to "go proxy," i.e., the set which is being reduced to $a$ and its extensional equivalents. We need the following preliminary definitions:

**DEF 4.14:** For all $a \in D$, the *type* $t(a)$ of $a$ is 0 if $a \in D_1$; otherwise, $t(a)$ is the $a$ such that $a \in L_\alpha^{ZF}$.

**DEF 4.15:** For all $A \in L^S$, $A^a$ is the result of replacing each variable $v_i$ in $A$ by $v_i^a$.

$A^a$ is, accordingly, the result of adding the superscript $^a$ to each of $A$'s variables. (In terms of the way our variables are defined, of course, this amounts to replacing some sets in a sequence by others.) Clearly $A^a$ is a formula of $L_\alpha^{ZF}$. The proxy function is now defined by induction on the type of members of $D$.

**DEF 4.16:** For all $a \in D$,

1. if $t(a) = 0$, then $\delta(a) = \emptyset$,
2. if $t(a) = a > 0$ (i.e., $a$ is a formula $A^a(f^1, x^a)$ with free variable $x^a$ and possibly constants $'c'$), then $\delta(a) = \{b \in L_\alpha \mid L_\alpha \models A(y, x)[\delta(c), b]\}$.

**THM 4.6:** For all $a > 0$ and $a \in D_\alpha$, $\delta(a) \in L_\alpha$. 
Proof: We proceed by induction on \(a\). Suppose \(a = 1\). Then if \(a \in D_a\), \(t(a) = 0\) and so \(\delta(a) = \emptyset\). But \(\emptyset \in L_1\).

Suppose the theorem holds for all \(\beta < a\), where \(a > 1\). Pick \(a \in D_a\). Then either \(a \in D_1\), in which case \(\delta(a) = \emptyset \in L_a\), or for some \(\beta < a\), \(a\) is a formula \(A^\beta(\{c\},x^\beta)\) with free variable \(x^\beta\) and possibly constants \(\{c\} \in D_\beta\). In the latter case, \(\delta(a) = \{b \in L_\beta \mid L_\beta \models A(y,x)[\delta(c),b]\}\). By the inductive hypothesis, \(\delta(c) \in L_\beta\), so \(\delta(a)\) is semidefinable in \(L_\beta\). Therefore \(\delta(a) \in L_{\beta+1} \subseteq L_a\).

**THM 4.7:** For all \(a > 0\) and \(b \in L_a\), there is some \(a \in D_a\) such that \(b = \delta(a)\).

**Proof:** We proceed by induction on \(a\). Suppose \(a = 1\) and \(b \in L_a\). Now \(L_1 = \{\emptyset\}\), so \(b = \emptyset\). Therefore, since \(D_1\) is nonempty, there exists \(a \in D_a\) such that \(b = \delta(a)\).

Suppose the theorem holds for all \(\beta < a\), where \(a > 1\). Pick \(b \in L_a\). Then for some \(\beta < a\), \(b = \{d \in L_\beta \mid L_\beta \models A(y,x)[c,d]\}\), where \(A(y,x) \in L^S\) and \(c \in L_\beta\). Now by the inductive hypothesis there are \(e \in D_\beta\) such that \(c = \delta(e)\). So \(\{e\}\) are constants of \(L^ZF\). Hence \(A^a(\{e\},x^a)\) is a formula of \(L^ZF\) and so is in \(D_a\). But \(b = \delta(A^a(\{e\},x^a))\).

**THM 4.8:** For all \(a > 0\) and \(a, b \in D_a\), \(\delta(a) \in \delta(b) \Rightarrow d \in a\) and \(\models_{ZF} b[d]\) for some \(d \in D_a\).

**Proof:** We proceed by induction on \(a\). Trivial if \(a = 1\). Suppose the theorem holds for all \(\beta < a\), where \(a > 1\). We
show that it holds for a.

CLAIM: For all \( \beta < \alpha \), \( A \in L^S \), and \( a \in D_\beta \),

\[ \models_{ZF} A^\beta[a] \iff L_\beta \models A[\delta(a)]. \]

Proof of the CLAIM is by the inductive hypothesis and induction on the construction of \( A \). Pick \( \beta < \alpha \) and \( A \in L^S \) and suppose that \( A \) is atomic (i.e., \( A \) is \( x \in y \) for variables \( x \) and \( y \)). Choose \( a \in D_\beta \).

Now \( \models_{ZF} A^\beta[a] \iff \models_{ZF} x^\beta \in y^\beta[a_1,a_2] = d \in a_1 \) and

\[ \models_{ZF} a_2[a] \text{ for some } d \in D_\beta \iff \text{(by the inductive hypothesis)} \]

\( \delta(a_1) \in \delta(a_2) \iff \text{(since by THM 4.6, } \delta(a_1), \delta(a_2) \in L_\beta \)

\( L_\beta \models x \in y[\delta(a_1), \delta(a_2)] = L_\beta \models A[\delta(a)]. \)

Suppose the CLAIM holds at \( \beta \) for all formulas of \( L^S \) shorter than \( A \). Then if \( A \) is \( \neg B \), \( \models_{ZF} A^\beta[a] \iff \models_{ZF} B^\beta[a] \)

\( \text{ (by the inductive hypothesis) } L_\beta \models B[\delta(a)] \iff L_\beta \models A[\delta(a)]. \) If \( A \) is \( (B \rightarrow C) \), the reasoning is similar. Now suppose \( A \) is \( \exists x B \). Then \( \models_{ZF} A^\beta[a] \iff \models_{ZF} B^\beta[a,b] \) for some

\( b \in D_\beta \iff \text{(by THM 4.6, THM 4.7 and the inductive hypothesis)} \)

\( L_\beta \models B[\delta(a), \delta(b)] \) for some \( b \in D_\beta \iff L_\beta \models A[\delta(a)]. \) This completes the proof of the CLAIM.

Returning now to the proof of the theorem, pick \( a, b \in D_\alpha \) and suppose \( \delta(a) \in \delta(b) \). Since \( \delta(b) \neq \emptyset \), \( b \notin D_1 \), and so for some \( \beta < \alpha \), \( b \) is a unary formula \( B^\beta(\langle c \rangle, x^\beta) \) with free variable \( x^\beta \) and possibly constants \( \langle c \rangle \) such that \( c \in D_\beta \). Hence \( \delta(b) = \{ d \in L_\beta \mid L_\beta \models B(y,x)[\delta(c),d] \} \). But since \( \delta(a) \in \delta(b) \), \( L_\beta \models B(y,x)[\delta(c), \delta(a)] \). So by the CLAIM \( \models_{ZF} B^\beta(y^\beta, x^\beta)[c,a] \). Therefore \( \models_{ZF} B^\beta(\langle c \rangle, x^\beta)[a] \), i.e.,
\[ \vdash_{ZF} b[a]. \] Hence, trivially, \( d \triangleq a \) and \( \vdash_{ZF} b[d] \) for some \( d \in D_a \).

On the other hand, suppose \( d \triangleq a \) and \( \vdash_{ZF} b[d] \) for some \( d \in D_a \). Since \( b \) is nonnull, again \( b \) is a formula \( B^\beta(\langle c \rangle, x^\beta) \), so \( \vdash_{ZF} B^\beta(y^\beta, x^\beta)[c, d] \). Hence by COROLLARY 4.2, \( \vdash_{ZF} B^\beta(y^\beta, x^\beta)[c, a] \). So by the CLAIM, \( L_\beta \models B(y, x)[\delta(c), \delta(a)] \). But then \( \delta(a) \in \{ d \in L_\beta | L_\beta \models B(y, x)[\delta(c), d] \} = \delta(b) \).

COROLLARY 4.9: For all \( a, b \in D \), \( \delta(a) \in \delta(b) \iff d \sim a \) and \( \vdash_{ZF} b[d] \) for some \( d \in D \).

**Proof:** Pick \( a, b \in D \). Then for some \( a, a, b \in D_a \). By THM 4.8, \( \delta(a) \in \delta(b) \iff d \sim a \) and \( \vdash_{ZF} b[d] \) for some \( d \in D_a \). Now since for all \( d \in D \) and \( c \in D - D_a \), \( \nmid_{ZF} d[c] \), it follows that for all \( c, d \in D_a \), \( c \triangleq d \iff c \sim d \). Hence \( \delta(a) \in \delta(b) \iff d \sim a \) and \( \vdash_{ZF} b[d] \) for some \( d \in D \).

COROLLARY 4.10: For all \( A \in L^S \) and \( a \in D \), \( \models_S A[a] \iff L \models A[\delta(a)] \).

**Proof:** By COROLLARY 4.9, DEF 4.10, and an induction on the construction of \( A \).

COROLLARY 4.11: \( \models_S ZF + V = L \).

**Proof:** Since \( L = ZF + V = L \), this follows by COROLLARY 4.10.
Since $V = L$ implies the axiom of choice and the generalized continuum hypothesis, these additional axioms are true on the standard interpretation of $L^S$ as well.

On the assumption of ZF's consistency, we have now shown that the representatives of formulas of $L^Z$ in our model form a model of ZF + $V = L$. Thus any hierarchy of possible tokens whose syntax and semantics are defined after the manner of Def 4.3 and Def 4.6 is also a model for ZF + $V = L$. In terms of the discussion of Chapter 2, these definitions constitute the grammatical rules for such a hierarchy.

Because of the power of ZF, these rules are and must be incomplete. Since, for example, the question of whether there is an inaccessible cardinal seems to be independent of ZF + $V = L$, they probably imply neither that there is nor that there is not a formula of $L^Z$ satisfying the formula of $L^S$ expressing "$x$ is an inaccessible cardinal."

This does not mean, of course, that the class $L^Z$ or the interpretation of $L^S$ over this class is incompletely defined. After we chose our arbitrary model of ZF, Def 4.3 picked out a definite subclass of this model and Def 4.6 provided a definite interpretation for the formulas of this subclass. But since the model chosen was arbitrary, it could have had any of a great variety of structures. Now we inferred from the fact that $L^Z$, defined with respect to this model, is a hierarchy of languages that tokens of a
hierarchy modelling $ZF + V = L$ are possible. But since we do not know the precise structure of the model, we know the structure of this hierarchy of possible tokens only within certain limits—limits drawn by the range of possible models of $ZF$. As suggested in Chapter 2, we may regard ourselves as having defined, not a single hierarchy of possible tokens, but a range of them—namely, all the hierarchies whose tokens obey the specifications laid down in DEF 4.3 and DEF 4.6. By COROLLARY 4.11, we know that each hierarchy in this range is a model of $ZF + V = L$.

The tokens of these hierarchies are very sophisticated bits of language. Many of them contain constants consisting of nondenumerably many symbols. Moreover, we have had to make substantial alterations in our original analysis to obtain a hierarchy of languages which models $ZF$. First, we have had to change our idea about the interpretation of $\epsilon$. Instead of satisfaction, it now signifies satisfaction by an extensional equivalent. Secondly, we thought initially that sets should be reducible to the unary formulas used to define them. But the formulas used to define the sets of $ZF$ are formulas of $L^S$, while those over which we have interpreted $L^S$ are formulas of $L^{ZF}$. This seems to reduce the credibility of the idea that the set theorist is talking about formulas. He works with formulas of $L^S$, so that it is reasonable to think that they are what he is talking about. But he has never heard of the formulas of $L^{ZF}$—which, after
all, is just a series of languages cooked up expressly for the purpose of modelling ZF. To what extent have these changes weakened the explanatory power of our analysis?

The shift in meaning of ∈ represents no radical departure. Our original idea was that sets should be regarded as open formulas viewed from a language in which extensionally equivalent formulas are indiscernible. Now when $L^S$ is interpreted over a single object language of a noncumulative hierarchy, extensionally equivalent formulas are indiscernible if ∈ is taken simply as a satisfaction predicate. With a cumulative hierarchy, however, extensionality can be preserved only if ∈ expresses satisfaction by an extensional equivalent. The extra complication is required by cumulativity and by the principle that extensional equivalents must be indiscernible if formulas are to "appear" as sets.

With regard to the objection that the formulas constituting set theory's subject matter are no longer the ones the set theorist actually uses, it should be noted, first of all, that we are not saying that $L^{ZF}$ is the hierarchy over which ZF quantifies. It is merely an example of the sort of hierarchy which models ZF. There is, in fact, no one definite hierarchy which can be regarded as the subject matter of ZF, as we shall see in Chapters 6 and 7. Therefore, it is not clear that ZF does not quantify over the formulas which set theorists actually use. We shall now
take a brief glance at one way in which it can be regarded as doing so.

Notice that, though $L^S$ and $L^{ZF}$ are different languages, the difference between them is actually quite small. Simply omitting constants and variable superscripts from any $L^{ZF}_\alpha$ gives $L^S$. In fact, instead of thinking of each $L^{ZF}_\alpha$ as a separate language, we can think of each as the same language --$L^S$ with added constants--continuously reinterpreted over itself to generate the hierarchy. Furthermore, the constants can be dispensed with, if we think of the domain as formulas of $L^S$ on various assignments to all but one of their free variables. That is, instead of formulas of the form $\forall A(\{c\},x)$, we might take as objects of quantification formulas of the form $\forall A(y,x)$ on the reading which assigns c to y, i.e., on the reading where the variables y function like names of the objects c.

Let us call a formula together with an intended reading for all but one of its free variables a saturated formula. The picture we arrive at, then, is as follows. We begin as before with a domain $D_1$ of arbitrary objects and interpret, not $L^{ZF}_1$, but $L^S$, over this domain. Then, instead of moving up to a new language $L^{ZF}_2$, we interpret $L^S$ again over its own saturated formulas as previously interpreted. This process is continued transfinitely. It is easy to see that the result is a structure precisely isomorphic to $L^{ZF}$ on its standard interpretation--a
structure which therefore models ZF + V = L, but whose components are just saturated formulas of $L^S$ interpreted over a sequence of ever-widening domains of saturated formulas of the same sort. The set theorist can thus be understood as quantifying over the formulas of $L^S$ on various reinterpretations. The formulas which he manipulates and the formulas he is talking about are therefore one and the same, though he thinks of them as interpreted differently: over the entire hierarchy if he is using them; over parts of it if he is mentioning them. This eliminates the need for an infinity of constants and an infinity of languages, though it requires instead that we think of the "construction" of the ZF universe as an unending process of reinterpretation. ZF is still about possible tokens, but now they are conceived as tokens whose interpretation shifts from level to level in this process.

This picture preserves the central and most important feature of our analysis: the idea that at least some of the objects which the set theorist is talking about are concrete, actual things which we can see, manipulate and learn about. In the next chapter, we shall see how a different process of reinterpretation may lead to an entirely different set theory.
Each of the set theories discussed thus far can be regarded as quantifying over a Tarski hierarchy or a language hierarchy of a closely related sort. These set theories and language hierarchies have a natural affinity; both avoid paradox by stratification of sets (formulas) into distinct orders, levels, or types (languages). However, some set theories, notably Quine's NF and ML, appear to avoid paradox without such stratification—indeed, without so much as a ban on self-membership.

This suggests that paradox can also be avoided on the semantic side without stratification into object language and metalanguage. It suggests, that is, that in some sense a language can contain its own satisfaction predicate. Let us say that a binary predicate P of language L on interpretation I is a satisfaction predicate for L if for all unary formulas A, B of L in the domain of I,

\[(1) \models_I Pxy[A,B] \iff \models_I B[A], \text{ and} \]
\[(2) \models_I \neg Pxy[A,B] \iff \models_I \neg B[A]. \]

Now Grelling's paradox shows that there is no classical
language $L$ and classical interpretation $I$ of $L$ meeting the following conditions:

(C1) $L$ has a satisfaction predicate $P$ on $I$,

(C2) $I$'s domain contains each unary formula of $L$,

(C3) $I$ is bivalent (i.e., for each formula $A$ of $L$ and variable assignment $f$ on $I$'s domain, $I(A,f) = T$ or $I(A,f) = F$).

For from (C1)-(C3) it follows that $\models_I \sim Pxx[\sim Pxx] = \not\models_I \sim Pxx[\sim Pxx]$. A classical language can contain its own satisfaction predicate only if (C2) or (C3) is false.

In "Outline of a Theory of Truth," Kripke (1975) shows that abandonment of (C3) alone is sufficient. In this chapter we shall sketch a Kripke-style theory of satisfaction and offer some speculations about its possible use in interpreting NF and related systems as theories for languages containing their own satisfaction predicates.

The nonbivalent semantics required by Kripke's theory can be developed in one of two ways--either by employing a "three-valued" logic after the manner of KLEENE (1962) (see pp. 332-40) or by using the supervaluation approach of VAN FRAASSEN (1966). We shall use what is essentially a variant of the former. The central idea of this approach is that interpretations may be partial functions. Given a formula and a variable assignment, an interpretation may assign the pair the value $T$, the value $F$,
or nothing at all—in which case the interpretation is undefined for that pair.³ This requires modification of the notion of a predicate assignment (see Chapter 1) to permit n-ary predicates to be true of some n-tuples, false of some and undefined for still others. Accordingly, we now redefine the notion of a predicate assignment on a domain D as an assignment to each n-ary predicate P^n of a pair <X,Y> such that X,Y ⊆ D^n and X ∩ Y = ∅. The first member of this pair is P^n's extension (the n-tuples of which it is true); the second its antiextension (the n-tuples of which it is false). Some n-tuples may belong to neither; these are the n-tuples for which P^n is undefined. We write \( \text{EXT}_p(p^n) \) and \( \text{ANTIEXT}_p(p^n) \) to signify P^n's extension and antiextension, respectively, on predicate assignment p.

A predicate assignment p on domain D induces an interpretation \( I_p \) on a classical language L as follows. For all formulas A of L and variable assignments f on D:

1. If A is \( P^n x_1 \ldots x_n \), then
   (i) \( I_p(A,f) = T \) if \( <f(x_1), \ldots, f(x_n)> \in \text{EXT}_p(p^n) \),
   (ii) \( I_p(A,f) = F \) if \( <f(x_1), \ldots, f(x_n)> \in \text{ANTIEXT}_p(p^n) \),
   (iii) \( I_p(A,f) \) is undefined otherwise;

2. If A is ~B, then
   (i) \( I_p(A,f) = T \) if \( I_p(B,f) = F \),
   (ii) \( I_p(A,f) = F \) if \( I_p(B,f) = T \),
(iii) $I_p(A, f)$ is undefined otherwise;

(3) if $A$ is $(B \lor C)$, then

(i) $I_p(A, f) = T$ if $I_p(B, f) = F$ or $I_p(C, f) = T$,
(ii) $I_p(A, f) = T$ if $I_p(B, f) = T$ and $I_p(C, f) = F$,
(iii) $I_p(A, f)$ is undefined otherwise;

(4) if $A$ is $\exists x B$, then

(i) $I_p(A, f) = T$ if $I_p(B, f^{d/x}) = T$ for some $d \in D$,
(ii) $I_p(A, f) = F$ if $I_p(B, f^{d/x}) = F$ for all $d \in D$,
(iii) $I_p(A, f)$ is undefined otherwise.

(Recall that $f^{d/x}$ is the function just like $f$ except that it assigns $d$ to $x$.) Clearly each predicate assignment $p$ induces one and only one interpretation $I_p$.

A predicate assignment $p'$ on language $L$ extends (or is an extension of) predicate assignment $p$ if for all predicates $P$ of $L$, $EXT_p(P) \subseteq EXT_{p'}(P)$ and $ANTIEXT_p(P) \subseteq ANTIEXT_{p'}(P)$. Likewise, $I_{p'}$ extends $I_p$ if $p'$ extends $p$. It is easy to show by induction on the construction of formulas that if $I_{p'}$ extends $I_p$, then $I_p \subseteq I_{p'}$, and all truths or falsehoods of $I_p$ are also truths or falsehoods of $I_{p'}$.

A predicate assignment $p$ on a domain $D$ is total for the predicate $P^n$ if $EXT_p(P^n) \cup ANTIEXT_p(P^n) = D^n$. A predicate assignment for $L$ is total if it is total for all predicates of $L$. An interpretation is total if it is induced by some total predicate assignment. A total interpretation
assigns either T or F to each pair consisting of a formula and variable assignment. That is, it is bivalent.

We shall now consider how the language $L_0$ of Chapter 3 can be interpreted by Kripke's method as a language which contains its own satisfaction predicate. (Recall that $L_0$ is the classical language whose only primitive predicates are $\epsilon$ and $P$. In what follows, the superscript '0' on $L_0$'s variables will play no role, since we shall not consider $L_0$ as part of a hierarchy. We therefore ignore the superscript.)

The interpretation of $L_0$ is constructed in stages by defining a series of interpretations $I_\alpha$ for $L_0$ over the domain $D = \{A \mid A$ is a unary formula of $L_0\}$. Each $I_\alpha$ is induced by a predicate assignment $p_\alpha$; and the $p_\alpha$ are such that if $\alpha < \beta$, then $p_\beta$ extends $p_\alpha$ (and hence $I_\beta$ extends $I_\alpha$).

We begin the interpretive process by setting $p_0(P) = <\emptyset, D>$ and $p_0(\epsilon) = <\emptyset, \emptyset>$. $P$ serves essentially the same function here as it did in Chapter 3. Regardless of whether the satisfaction predicate occurs in a metalanguage or in its own object language, its interpretation cannot begin until we have some formulas whose satisfaction conditions are already defined. $P$ provides such formulas.

Clearly any formula not containing $\epsilon$ will be either true or false on any variable assignment at $I_0$ (i.e., $I_{p_0}$). But formulas containing $\epsilon$ will not in general have truthvalues. $\epsilon$ is not a satisfaction predicate under $I_0$ since, for example, $\models_{I_0} \neg P[x[Px]]$ but $\not\models_{I_0} x \epsilon y[Px, \neg Px]$. 
Given an interpretation $I_\alpha$, we construct the predicate assignment $p_{\alpha+1}$ as follows:

(i) $\langle A,B \rangle \in \text{EXT}_{p_{\alpha+1}}(\epsilon) \iff I_\alpha \models B[A],$

(ii) $\langle A,B \rangle \in \text{ANTIEXT}_{p_{\alpha+1}}(\epsilon) \iff I_\alpha \models B[A].$

(The interpretation of $P$ remains the same, i.e., $p_\alpha(P) = \langle \emptyset,D \rangle$ for all $\alpha$.) The assignment $p_{\alpha+1}$ then induces $I_{\alpha+1}$ as described above. For limit ordinals $\alpha$, we let

(i) $\langle A,B \rangle \in \text{EXT}_{p_{\alpha}}(\epsilon) \iff \langle A,B \rangle \in \text{EXT}_{p_{\beta}}(\epsilon)$ for some $\beta < \alpha,$

(ii) $\langle A,B \rangle \in \text{ANTIEXT}_{p_{\alpha}}(\epsilon) \iff \langle A,B \rangle \in \text{ANTIEXT}_{p_{\beta}}(\epsilon)$ for some $\beta < \alpha.$

Each succeeding interpretation extends all previous ones. As the interpretive process continues, the extension and antiextension of $\epsilon$ get larger, and satisfaction is defined for more and more pairs of unary formulas. Kripke shows that at some transfinite level $\alpha$ the process reaches a saturation point at which growth ceases, so that $I_\beta = I_\alpha$ for all $\beta > \alpha.$ At such a fixed point, as Kripke calls it, $\epsilon$ becomes a satisfaction predicate. (Otherwise, there would be formulas $A$, $B$ such that $I_\alpha \models B[A]$ but $\langle A,B \rangle \notin \text{EXT}_{p_{\alpha}}(\epsilon)$ or $I_\alpha \models B[A]$ but $\langle A,B \rangle \notin \text{ANTIEXT}_{p_{\alpha}}(\epsilon),$ and $\epsilon'$s extension or antiextension would continue to grow.) For some pairs $\epsilon$ remains undefined, but the satisfaction conditions for these pairs are also undefined.
Among the formulas whose satisfaction conditions are not fully defined at the fixed point are all those which generate paradox (e.g., $\neg x \in x$) and all those which do not generate paradox, but which seem nevertheless to lack complete satisfaction conditions. An instance of the latter type is $x \in x$, which intuitively is neither true nor false of itself. Kripke calls such formulas ungrounded.

Kripke shows that if we begin the interpretation, not by assigning $<0,0>$ to $\in$, but by assigning arbitrary satisfaction conditions to ungrounded but unparadoxical formulas like $x \in x$, we ultimately arrive at a new fixed point. (The fixed point attained by assigning $<0,0>$ to $\in$ is called the minimal fixed point, because interpretations defined at all other fixed points extend the interpretation defined at it.) At this new fixed point, $\in$ will once again be a satisfaction predicate, but some formulas will have more fully defined satisfaction conditions than before. Some formulas, however,--specifically the paradoxical ones, like $\neg x \in x$--lack complete satisfaction conditions at all fixed points. There is no fixed point, for example, at which $\neg x \in x$ is either true or false of itself.

Loosely speaking, Kripke's theory can be seen as a way of collapsing Tarski hierarchies down into their base languages. Type restrictions vanish, and the only remaining vestige of the hierarchical structure is the series of interpretations by which the satisfaction predicate
is defined.

Intriguingly, Quine's NF does roughly the same thing for the simple theory of types (STT). STT is the theory axiomatized by $\text{EXT}(S)$ and a comprehension schema like $\text{COMP}$ of Chapter 3, except that Clause (3) is omitted. NF is syntactically like STT, except that variable types are omitted. The language of NF is thus the standard language $L^S$ of set theory. Its axioms are $\text{EXT}$ and the following comprehension schema:

$\text{COMP}(\text{NF})$: $\exists y \forall x (x \in y \leftrightarrow A)$, where (1) $x$ is free in $A$, (2) $y$ is not free in $A$, and (3) $A$ is stratified.

This last condition, that $A$ be stratified, is simply the requirement that $A$ would be well-formed in the language $L$ of type theory if variable superscripts were restored. More precisely, a formula $A$ of $L^S$ is stratified if there is an assignment $S$ of natural number superscripts to the variables of $A$ such that all occurrences of the same variable have the same superscript, and for each subformula of the form $x \in y$, $S(y) = S(x) + 1$.

The universe of NF exhibits no types or levels, and NF's variables range over all its members without restriction. Thus NF collapses the type hierarchy of STT into a single level or layer, much as Kripke's theory collapses a Tarski hierarchy into a single language.
STT is equivalent (allowing for notational differences and differences in the treatment of relations) to Russell's ramified theory of types with the axiom of reducibility.\textsuperscript{5} We saw in Chapter 4 that Russell's theory can readily be understood as a Tarski hierarchy only so long as the axiom of reducibility is not added. Thus we should expect there to be no straightforward way of regarding STT as quantifying over a Tarski hierarchy; and, indeed, I know of no way of doing so.

The type theory most easily interpretable over a Tarski hierarchy is CTT, as we saw in Chapter 3. Now CTT is related to a fragment of NF, which we shall call NFP, in the same way that STT is related to NF.\textsuperscript{6} That is, speaking picturesquely, NFP collapses CTT's universe into a single level. But then, since CTT is modelled by the predicative unary formulas of a Tarski hierarchy, and a Kripke-style theory of satisfaction is just the "collapse" of a Tarski hierarchy, it seems reasonable to think that NFP is modelled by a corresponding subclass of the formulas of the language of a Kripke-style satisfaction theory. I have not been able to show this, but it is rather easy to prove certain partial results, which are detailed below.

NFP is in some ways more powerful than CTT (it contains the axiom of infinity), but less powerful than NF (it does not permit proof of the least upper bound theorem). Its axioms are EXT and the following comprehension schema:
COMP(NFP): $\exists y\forall x(x \in y \leftrightarrow A)$, where (1) $x$ is free in $A$, (2) $y$ is not free in $A$, and (3) $A$ is predicative on $x$.

A formula $A$ of $L^S$ with $x$ free is **predicative on $x$** if there is an assignment $S$ of natural number superscripts to the variables of $A$ such that:

1. all occurrences of the same variable have the same superscript,
2. for each subformula of the form $z \in y$, $S(y) = S(z) + 1$,
3. for no variable $v$ is $S(v) > S(x)$,
4. for no bound variable $v$ is $S(v) \geq S(x)$.

(Clearly each formula predicative on $x$ is also stratified, so that NFP is a subtheory of NF.) It will be useful to extend the definition of predicativity on a variable to the language $L_0$ (which is like $L^S$, except that it contains the additional predicate $P$) by adding a fifth clause:

5. for each subformula of the form $Px$, $S(x) = 0$.

We interpreted CTT over the unary predicative formulas of $L$. We should like some way to characterize the analogous class of formulas of $L^S$ or $L_0$. Let us say that a unary formula of $L^S$ or $L_0$ is **predicative if it is predicative on its free variable**. A formula of $L^S$ or $L_0$ is thus predicative...
if it becomes a predicative formula of $L$ with the addition of superscripts to its variables.

Our conjecture is that NFP is modelled by the unary predicative formulas of a Kripke-style theory of satisfaction. But all Kripke-style theories have truth-value gaps. Must not the interpretation we give to NFP, then, have truth-value gaps as well?

Not necessarily. NFP is to be interpreted only over predicative formulas (of $L_0$). Predicative formulas, like stratified formulas generally, seem not to give rise to paradox. (After all, no contradiction has ever emerged in NF.) Since only paradoxical formulas remain without truth-value at all fixed points, it seems likely that by interpreting our Kripke-style theory at each level only over predicative formulas of $L_0$, we could arrive at a fixed point at which each pair of predicative formulas is in either the extension or the antecedent of $\epsilon$. Let us call such a fixed point predicative-saturated or p-saturated. The interpretation at a p-saturated fixed point would be bivalent, since variables would range only over predicative formulas and both $P$ and $\epsilon$ would be defined for all predicative formulas. Thus at a p-saturated fixed point, we trade the falsity of (C3) for the falsity of (C2). Bivalence is restored at the cost of expressing the satisfaction conditions only of a proper subclass of $L_0$'s formulas. Moreover, at a p-saturated fixed point NFP would be true.
THM 5.1: Let $I$ be the interpretation defined at a $p$-saturated fixed point. Then $\models_I \text{EXT}$.

Proof: The reasoning is entirely analogous to that of the proof of THM 3.1, except that reference to variable superscripts and distinct domains is dropped.

THM 5.2: Let $I$ be the interpretation defined at a $p$-saturated fixed point. Then if $B(z)$ is an instance of $\text{COMP}(\text{NFP})$, $\models_I B(z)[a]$ for all predicative formulas $a$ of $L_0$.

Proof: Let $B(z)$ be an instance of $\text{COMP}(\text{NFP})$. Then $B(z) = \exists y \forall x (x \in y \leftrightarrow A(z,x))$, where $A$ is a formula of $L^S$, $x$ but not $y$ is free in $A$, and $A$ is predicative on $x$. Pick predicative formulas $a$ of $L_0$ and let $C(x)$ be the result of taking the $z_i$-elimination of $A(z,x)$ for each variable $z_i$ among the $z$ with respect to the corresponding object $a_i$ among the $a$. (See DEF 3.8.) $C(x)$ is a predicative formula of $L_0$. Now by reasoning similar to that used in the proof of LEMMA 3.3, $\models_I C(x)[b] = \models_I A(z,x)[a,b]$ for all predicative formulas $b$ of $L_0$. Hence $\models_I \forall x (x \in y \leftrightarrow A(z,x))[C(x),a]$, and so $\models_I B(z)[a]$.

THM 5.1 and THM 5.2 show that if a $p$-saturated fixed point exists, then NFP is modelled by the predicative formulas of a Kripke-style theory of satisfaction.

But what of NF itself? Just as NFP might be modelled by the predicative formulas of a Kripke-style theory of satisfaction, might
not NF be modelled by the stratified ones? If instead of defining successive interpretations only over predicative formulas, we chose to define them over all stratified formulas, might we not reach a fixed point at which satisfaction is defined for each pair? Again, this seems reasonable, since stratified formulas are apparently not paradoxical. Let us say that such a fixed point is stratified-saturated or s-saturated. Unfortunately, the reasoning of THM 5.1 and THM 5.2, which depends upon predicativity, breaks down for s-saturated fixed points, so that the obvious way of showing that such a fixed point provides a model of NF is blocked. And I have not been able to find an alternative method of proof. 7

A great deal of work remains to be done here. We need to know whether p-saturated or s-saturated fixed points exist and whether the latter, if they exist, provide models for NF. Positive answers to these questions would reveal close ties between NF and Kripke's theory. A positive answer to the second one would show that NF is consistent. In either case, positive answers would show that our analysis can be extended in a very natural way to systems like NF and NFP.

Having outlined the details of our analysis for a number of different set theories, we are at last ready to return to philosophical matters.
In Chapter 2 we claimed that analyzing sets as open formulas permits a cogent account of set-theoretic knowledge. It is now time to make good on that claim. In the present chapter we consider several closely related questions about the set theorist's knowledge. We begin with the question of how students of set theory learn the subject. This leads us to a discussion of how the founders of set theory obtained their knowledge. Our conclusion will be that they did not, strictly speaking, obtain it; they created it. To invent a set theory is to establish grammatical conventions for a language or hierarchy of languages. Since languages, on our view, consist of possible formula tokens, this leads us finally to the question of how knowledge of nonactual tokens is possible.

The question with which we begin is a simple one, and the answer it calls for is neither difficult nor profound.

We simply want to know what the student of set theory is taught and by what means he is taught it. To find out, we turn to the classroom.
In elementary set theory or mathematics classes, sets are almost invariably introduced as groups or collections of things. As examples, the instructor may list such objects as The United States Senate, a flock of birds, a bunch of grapes, or a set of dishes. (These are actual examples culled from college mathematics texts.) Thus is initiated the confusion, much lamented in Chapter 1, between sets and collections.

When actual work begins, however, talk of such collections is soon forgotten. Most students begin with naive set theory. Certain apparent sets, they are told, are paradoxical; a few examples, such as the Russell set, are discussed, and then attention is gently diverted from these "pathological cases." It is noted that for most practical applications one can assume that every open formula defines a set, and work proceeds on this assumption.

What the student is actually taught to do is to solve problems, i.e., to manipulate formulas in certain well-recognized ways. It is no part of set-theoretic training to make observations of The United States Senate, flocks of birds, bunches of grapes, or sets of dishes. When a student is asked to prove the existence of a set, that request is considered fulfilled if he produces an appropriate open formula in curly brackets. Moreover, if he tried to go beyond this, if he tried, for example, to give evidence in his proof that the world actually contains a collection
defined by the formula he produced, his behavior would be considered inappropriate.

The intermediate student progresses from naive set theory to axiomatics. Here the paradoxes are taken seriously, and it is no longer sufficient in proving the existence of a set merely to give a formula expressing its membership conditions. Rigorous deduction from the axioms is required. But that is all that is required; the axioms themselves are taken on faith.

Set theory is peculiar in this regard. Even in the highest reaches of theoretical physics, if the instructor makes a claim of which his students are sceptical, the standard response is to point to evidence and data and to explain how these support the claim. But when the set theorist introduces his axioms, if students are sceptical of their truth, there is no standard response. The set theorist may resort to any (or all) of the following ploys, some of which are mutually inconsistent: "You can't prove anything without assumptions, and these are the most fruitful assumptions we know," or "We don't care whether the axioms are true or false; we are simply interested in their consequences—what things would be like if they were true," or "These axioms define what it is to be a set; anything which obeys the axioms is a set, and so they are true of sets by definition," or "These axioms represent our best guesses about the behavior of collections; they satisfy our
intuitions about how collections ought to behave" (this ploy is often accompanied by an appeal to the "naturalness" and "inevitability" of each of the axioms), or "These axioms have been inferred by working backwards from know mathematical results and trying to find foundations for them," or "Just accept them for now; after you get used to them their truth will become obvious to you." I have encountered versions of each of these gambits in textbooks or in classrooms. Their lack of unanimity arouses suspicion that often instructors are simply giving rationalizations to silence objections, so that work can begin on the material of "real interest." And in practice such objections are usually effectively silenced without presentation of anything remotely like solid evidence for the axioms' truth; students invariably wind up accepting the axioms on authority. This, as we shall argue shortly, is no mere sociological accident.

First, however, some further observations about set-theoretic education are in order. Notice that it is never part of such an education to practice techniques of introspection or meditation to "sharpen the mind's eye." The successful student does not spend long hours in rapt contemplation of the cumulative hierarchy. Students learn, pencil in hand, by solving problems. Of course, practice in problem solving eventually results in the acquisition of certain "mental pictures" or intuitions which the student uses as heuristic guides. But he does not and could not
learn the facts of set theory by studying these pictures or intuitions. Initially, he has neither.

He may, of course, have mental pictures of ordinary collections; but these, as we have seen, are not pictures of sets. Indeed, they will conflict with what he is told about sets, and it is just this sort of conflict which causes many students to find the concept of a null set, for example, so exasperatingly counterintuitive.

He acquires pictures of and intuitions about sets, their relations to one another, and their arrangement in a cumulative hierarchy only gradually—in part, perhaps, by observing diagrams his instructor puts on the board, but also by trial and error in the process of problem-solving. His first mental pictures of sets are likely to be misleading; attending to them may cause him to waste fruitless hours on a simple problem. Finally, however, when a solution comes, or when it is shown to him, he understands why the pictures were unhelpful, and he revises them so that they will be more efficient problem-solving aids in the future. In this way he constantly expands and changes his mental pictures in response to his problem-solving needs. What he learns he learns by doing problems and by listening to his instructor; his mental pictures and intuitions codify, but do not add to, his knowledge. They do not play the role of perceptions. By codifying information efficiently and in an easily-accessed way, they may help to suggest new information
or solutions to problems. But this information is never regarded as valid unless it is checks out on paper. Thus the function of these mental goings-on is entirely subsidiary to the task of solving problems, i.e., manipulating formulas. And this is really the only thing the student of set theory is taught to do.

Having outlined some of the salient features of a normal set-theoretic education, we may now ask: What is it that the student, thus educated, has come to know? He has certainly acquired an ability to perform certain operations with formulas. But has he really come to know anything about sets as such? The examples of "sets" he was given at the beginning of his education were not sets at all, but collections. He has never been shown an actual set. At first, when he or his instructor "proved the existence" of a set, all that was done was to produce a defining formula. Later, in the axiomatic stage, "proof" of the existence of a set consisted of deduction of an existential statement from certain axioms, whose truth was never demonstrated to the student and which he accepted on the authority of his instructor. Now perhaps the instructor himself or his instructor had some way of knowing that these axioms were true of platonic sets and the student was justified in believing things about sets on the basis of this authority. (We shall consider this possibility in a moment.) But if this is not the case, then the student—who represents most
members of the philosophical, scientific, and mathematical communities—apparently has no justification for his belief in platonic sets at all. He has spent his entire set-theoretic education learning about and manipulating formulas.

In most cases, the instructor from who a student learns set theory learned it himself in a similar way. He, too, accepted the axioms on the authority of his instructor and never gave a great deal of thought to the problem of their truth. Now if anyone anywhere knows that there are real sets of which the axioms are true, then this chain of authority must have an end. There must have been an original authority who did not obtain his knowledge from someone else. Someone was the first to discover sets. But how and when did this discovery occur?

It is obvious that this question is wrong-headed. There was no such discovery. The earliest set theorists—Cantor and his immediate predecessors and successors—simply assumed as a matter of course that for each open sentence there was a corresponding set. (This was the inconsistent comprehension principle which Cantor used in much of his reasoning, though he did not explicitly formulate it.1) Like the naive set theorist of today, if asked to show the existence of a specific set, for the most part they simply produced a defining formula. They neither sought nor offered further evidence for their existential claims.
But was this all the evidence they had? Did they not also have mental pictures of sets or intuitions about them which constituted evidence? To think this is, as we argued above, a mistake. Mental pictures and intuitions only codify information obtained from other sources—information which may well be inaccurate. In the absence of assurances of the reliability of these other sources, our set-theoretic "pictures" or intuitions provide no more evidence for the existence of sets than our "pictures" of or intuitions about goblins provide for the existence of goblins.

The fact remains that the only substantial evidence these early set theorists actually gave—and hence the only substantial evidence we have reason to think they had—was the production of open formulas. But the production of an open formula is sufficient evidence only for belief in the existence (or possibility) of the formula (and, perhaps, in the existence of its producer). Therefore, either these early set theorists did not know that sets exist (since their evidence was insufficient), or knowing that sets exist is nothing more than knowing that certain formula tokens exist or are possible.

Succeeding set theorists had, if anything, even less reason to believe in platonic sets than Cantor did. Russell's antinomy had shown that Cantor's original intuitions about sets were wrong. New intuitions were not long in coming, but in a field where intuitions had failed so spectacularly,
their veracity was open to serious doubt. Yet Cantor's successors did not in general seek any evidence for the existence of sets beyond that bequeathed to them by Cantor. Like Cantor, they took the existence of at least some sets for granted. They were for the most part mathematicians rather than philosophers; and they saw their task, not as providing ontological or epistemological underpinnings for set theory, but as "trimming down" and reorganizing Cantor's universe to the point of consistency. (That they viewed their task this way is an important datum to which we shall return shortly.)

Our discussion therefore points to the conclusion that both Cantor and his successors lacked sufficient evidence to justify belief in platonic sets. But to some this conclusion will seem precipitous. It is true that no set theorist has ever sought or found observational confirmation of the existence of sets, but perhaps some have possessed broadly theoretical or pragmatic evidence. Perhaps set theorists have been justified in believing in sets by the fruitfulness of set theory as a source of important theorems and powerful mathematical techniques. And perhaps they have come to know of the existence of sets in this way.

But this is not the case. The utility of mathematical results and techniques which issue from set theory does not hinge on the ontological character of the objects (if any) of which the axioms are true. If, as we have demonstrated,
the axioms apply equally well to possible expression tokens as to sets, then their utility—though it may be evidence of their truth—is not evidence sufficient fully to determine the character of the objects of which they are true. Mathematical success gives no more evidence for the existence of platonic sets than it does for the existence of an infinite universe of actual expression tokens, or any other sort of object which could model the axioms. Decision about the ontological character of the objects of which the axioms are true requires collateral evidence—including consideration of which of the competing views provides the most elegant overall ontology, which accounts best for our knowledge of these objects, and which meshes most smoothly with our best scientific picture of the world. Since set theorists have not bothered to sort through such considerations, it follows that even if they had sufficient pragmatic evidence to know that their axioms were true, they have not had pragmatic evidence sufficient to justify the conclusion that their were true of platonic sets.

Our conclusion, therefore, stands. If set theory is about platonic sets, then its founders did not have adequate evidence to justify belief in them. Therefore, as we suggested earlier, modern students do not know that sets exist on the basis of their authority, since evidence adequate to justify belief in sets as such is not to be found anywhere along the line. Accordingly, it seems, no one knows that
sets exist.

We inferred this sceptical conclusion from the assumption that set theory is about platonic sets. It is a discomforting conclusion, but perhaps one with which the platonist could live. We did not conclude (here at least) that knowledge of sets as such is impossible. But we argued that consideration of the history of set theory, of the way it is taught, and of the fact that no one to date has produced adequate pragmatic evidence of the existence of platonic sets imply that no one actually does know that sets exist.

To widen our perspective for a moment, this entire study constitutes an argument that such pragmatic justification will not be found. The challenge it raises to the platonist is to show that positing of platonic sets provides a better picture of set theory, all things considered, than does an analysis of sets as possible open formula tokens or objects of some other sort. If the platonist can meet this challenge, then he will have good pragmatic evidence for his view. But, beginning with Chapter 1, we have been compiling difficulties for the platonistic view which make it seem unlikely that the challenge will be met.

That compilation is by no means complete. We now turn to some additional difficulties for the platonistic view.
On the assumption that set theory is about platonic sets, it is very difficult to make sense of the developments which followed the discovery of Russell's paradox. If the axiomatizers of set theory--Zermelo, Russell, von Neumann, and others--were engaged in scientific theorizing aimed at revealing facts about the platonic universe, then the methods they employed seem curiously inappropriate.

For one thing, they proceeded in apparent violation of Occam's razor. There has always been a tendency among set theoreticians to assume that the set-theoretic universe is as large as possible within the limits of consistency. The guiding idea has been to prune from Cantor's universe only as much as is necessary to excise the paradoxes. Accordingly, there has been little enthusiasm for axioms limiting the "size" of the universe--even when these have mathematically important consequences. But this seems to contrast sharply with general scientific practice. No physicist, for example, posits kinds or numbers of particles beyond those needed to account for his observational data. If the axiomatizers were formulating theories about a universe of actual objects, we should expect them to exhibit similar restraint, i.e., to posit only those entities whose existence is needed to explain whatever data it is that they are attempting to explain. (Incidently, saying what the data of set theory are and how they constrain its formulation are no easy tasks for the platonist.)
One might argue after the fashion of Lewis that keeping down the number of sets is spurious parsimony, since it is quantitative rather than qualitative. But this distinction is vague. A set theory which allows any new sets also allows new kinds of sets (e.g., inaccessible cardinals, Mahlo cardinals, etc.) with properties not displayed by sets of a more parsimonious theory. Thus the apparent ontological extravagance almost universally exhibited by set theorists is in some measure both quantitative and qualitative, and it is a curiosity which resists explanation on the platonistic view.

A further difficulty is that platonism renders the structure of the set-theoretic universe and the fundamental principles governing it ultimately inexplicable. Consider, for example, the idea that sets are arranged in a cumulative hierarchy. Why should the platonic universe be arranged this way? One way to answer is that this arrangement captures perfectly the intuitions surrounding a certain concept of set (what Boolos [1971] calls the "iterative" or "stage" concept). But this merely pushes the question back a step. Why should the platonic universe conform to this concept? If indeed the universe has the structure the concept ascribes to it, then why does it have this structure and not some other—say, the structure described by NF?

The platonist may reply that this is the structure produced by the iterative process of generating and
collecting "collections"—but then the question becomes: Why should we expect the set-theoretic universe to be the outcome of such a process? And what exactly is the significance of the "process" metaphor here? Why is it so curiously appropriate? The genuine platonist can hardly assert that the set-theoretic universe was actually created by such a process, for then he would face the question of when and by what forces the process occurred.

On the platonistic view, in other words, the arrangement of sets in a cumulative hierarchy is apparently just a brute fact, incapable of explanation. The universe of sets simply is as it is.

Similar points can be made with respect to other fundamental principles of various approaches to set theory—particularly Zermelo's limitation of size doctrine and Russell's vicious circle principle. If the platonic universe obeys these principles, why does it obey them? The platonist cannot explain.

Accordingly, if we regard set theory platonistically, then in addition to the metaphysical and epistemological problems catalogued in earlier chapters, we face the following difficulties: (i) no one knows that sets exist, (ii) set theory was developed in seeming violation of Occam's razor, and (iii) the fundamental cannons of set theory seem to be brute facts, incapable of explanation. These further difficulties, like the earlier ones, are all in some measure
alleviated by our analysis.

If we analyze sets as open sentences, then everyone who studies set theory has ample justification for belief in at least some of them, because he has perceived some. Hence it is easy to know that sets exist (though what this means on our analysis is that unary formulas are possible). To see how the analysis helps with problems (ii) and (iii), it will be helpful to retrace some of the major developments in set theory as they appear from its perspective.

The analysis makes it easy to understand the apparent self-evidence of Cantor's naive comprehension principle. To each unary formula there really does correspond a set (i.e., that formula itself and its extensional equivalents viewed indiscernibly). That much is a trivial consequence of the analysis. The principle's inconsistency arises, not from the assumption that to each formula there corresponds a set, but from the additional assumption implicit in the principle's use that satisfaction predicates can occur indiscriminately without regard for language levels. The early set theorists thought of all sets (formulas) as belonging to the same level (language); they had no notion of stratification into a hierarchy which prevents set-theoretic (semantic) paradox. In our terms, they thought of all formulas as belonging to a single language—a language which quantified over its own unary formulas and contained its own satisfaction predicate. But such a language is, as we now know,
impossible, without resort to truth-value gaps of a Kripke-
style theory of truth.

Appearance of the antinomies was an indication that
the current notion of satisfaction was incoherent. The work
of the succeeding axiomatizers--two paradigm examples being
Zermelo and Russell--now appears as an attempt to reformu-
late the conventions governing the satisfaction predicate
(or something akin to it) in a nonparadoxical way. To
accomplish this, each resorted to the device Tarski intro-
duced explicitly for the same purpose several decades
later--a hierarchy of languages. Russell, as we saw in
Chapter 4, set up what was essentially a full-blown Tarski
hierarchy, before he complicated it by adding the axiom of
reducibility.

Zermelo's system, on the other hand, provides only
the outlines, as it were, of a language hierarchy. His
axioms assert the possibility of a great many formula tokens
and tell us much about their satisfaction conditions, but
they imply almost nothing about the syntax of these formulas
and leave a great many semantic details undetermined. We
can think of Zermelo's axioms--as well as the axioms of
other set theories--as "very" incomplete grammatical rules,
rules which define the main features of a range of languages
or language hierarchies but leave many details out of
account.
One of the points of Chapter 4 was to show how such details could be filled in—at least in part—for the sort of language described by the axioms of ZF. Definitions 4.3 and 4.6 provided more explicit grammatical rules than do the axioms, even though they remain incomplete, as we noted there. (Of course, our way of filling in the details was to some degree arbitrary, and there are many other ways of doing it.)

By now an objection to this way of characterizing set theory has probably occurred to the reader. If what Zermelo, Russell, and the other axiomatizers were doing in effect was stipulating the conventions governing language hierarchies, why did they not see it that way? The objection requires several responses.

Russell, in fact, did see it that way, at least some of the time. He often spoke as if his propositional functions were open formulas, and he thought of the ramified theory of types as a semantic theory which embodied solutions to all the paradoxes—semantic as well as logical. But Russell also sometimes thought of propositional functions as something akin to properties, and Zermelo and many others certainly did not think of themselves as talking about language at all.

There are at least three reasons for this. First, Zermelo and most other set theorists confused sets with collections. (Russell partially disabused himself of this
confusion by regarding sets as "logical fictions" in a manner which in some ways foreshadows our treatment by identification of indiscernibles. Since collections are obviously not open formulas, most set theorists simply failed to make the connection. It never occurred to them that what they said about sets could be construed as talk about language.

The second reason is that even if it had occurred to them, to make sense of the idea they would have to have seen how open formulas can appear to be sets from the point of view of an abstract language; and there has been little or no discussion of abstract languages until comparatively recently.

And, third, even if they had seen this, they would perhaps have been stopped by the problem of there being insufficiently many formula tokens to make the reduction work. Modal quantification was also little discussed until recently. Thus the axiomatizers remained under what to us is the illusion of seeing possible concreta as actual abstracta—an illusion fostered by the abstraction and modality of the language in which they were working.

It can now be seen that our interpretation of the development of set theory avoids the remaining problems—(ii) and (iii)—raised against the platonic view. The axiomatizers did not violate Occam's razor, because they were not engaged in an investigation of reality. They were not
positing universes; they were creating languages through implicit stipulation of their grammatical rules. The attempt to salvage as many of Cantor's sets as possible was not a perversion of scientific method, but an attempt to make these languages as richly expressive as consistency permitted. What has often been viewed as a "thinning out" and reorganizing of Cantor's universe was instead a thinning out and reorganizing of the language of his theory of satisfaction.

This was no ontological enterprise. The axiomatizers were engaged in revising the syntax and semantics of the formal theory of satisfaction—splitting up the language in which it was expressed into an entire hierarchy of languages. Occam's razor does not apply to such a project, for the project itself involves no claims of existence and has nothing to do with what there actually is.6

Moreover, we need not regard the fundamental canons of various set theories as brute fact. The idea that sets are arranged in a cumulative hierarchy, the limitation of size doctrine, and the vicious circle principle all have explanations on our view. Each issues from conventions adopted after the discovery of Russell's paradox for the consistent application of the satisfaction predicate.

"Sets" (i.e., the unary formulas people take for sets) are arranged hierarchically, because the axiomatizers of set theory in effect fragmented Cantor's original
language and reorganized the fragments hierarchically in order to arrive at an apparently consistent, though incomplete, set of conventions governing the use of the satisfaction predicate. The metaphor of iterative construction—whose appropriateness in describing the hierarchy is puzzling on the platonistic view—is quite natural on our analysis; it appropriately characterizes the recursive process by which the languages of the hierarchy are defined.

The limitation of size doctrine becomes, on analysis, the principle that no formula is satisfied by as many formulas as the hierarchy itself contains. For transfinite hierarchies like $L^{\aleph_1}$, which are "larger" than any given cardinality, this is a direct consequence of the fact that each formula is satisfied only by formulas of lower levels, so that each is satisfied by at most as many formulas as occur at lower levels.

Likewise, Russell's vicious circle principle—the idea that "no totality can contain members defined in terms of itself"—is an expression of the convention that in a language hierarchy no formula $A$ is satisfied by formulas which $A$ satisfies, i.e., formulas whose satisfaction conditions are defined over a domain containing $A$.

All three principles are thus explicable as consequences of the stratification of languages adopted to ensure consistency in the theory of satisfaction. The platonist, as we saw, seems to have no explanation for them at all.
Because the axiomatizers of set theory were engaged in revising and redefining the language of the theory of satisfaction, rather than theorizing about the world, it is unsurprising that they sought no special evidence for the truth of their axioms. Evidence is not required for the stipulation of grammatical conventions.  

For the same reason, it is by no means peculiar—indeed, it is entirely appropriate—that students today are taught to accept the axioms of set theory on authority. The axioms are grammatical stipulations about the languages over which they quantify, and grammatical stipulations can be transmitted only by authority, never shown to be true.  

In summary, then, our picture is this. The reason students learn set theory exclusively by manipulating formulas is that set theory is about formulas—specifically, the possible formula tokens of a theory of satisfaction (or satisfaction by an extensional equivalent). The axioms of set theory are grammatical stipulations about these formulas and, accordingly, no evidence can or should be offered for their truth. The founders of set theory did not obtain their knowledge by investigation of a platonic universe; they created it by issuing these stipulations, or deduced it from the stipulations once they were issued.  

Of course, none of the founders of set theory thought of themselves as doing exactly this, for reasons we have mentioned. Our claim is not that this is what they intended
to do, but that it is what they are best understood as having done. To lay down set-theoretic axioms is to define (albeit incompletely) a language or hierarchy of languages, and to learn set theory is to learn about possible formula tokens of such a language or languages.

This last point brings us to an important issue. The possible tokens with which set theory is concerned are, many of them, nonactual; they have never existed and never will. We claimed in Chapter 2 that knowledge of nonactual expression tokens of various languages is commonplace. We now need to look more closely at the sources of that knowledge.

One source lies in our powers of construction. Though we have no direct sensory access to nonactual tokens, it does not follow that we need a metaphysical telescope or special inner eye to find out about them. We are, after all, capable of constructing them (and hence rendering them actual) with an apparatus as mundane as a pencil and paper. When a question about a specific sort of token arises, this constructional capacity often enables us to gain sensory access to a token of the type in question.

Still, our powers of construction are limited in ways in which our knowledge of possible tokens is not. We cannot in practice construct very large finite tokens—let alone infinite ones—, nor can we construct tokens of every type. Yet we know specific facts about large or infinite tokens and general facts about all possible ones. Whence
this superfluity of knowledge? The answer, I think, is fairly obvious.

A person learning a language begins by mastering the use of a finite number of simple words and phrases. For this it is clear that he needs to have encountered only finitely many actual tokens. Later, he gradually becomes aware of the patterns according to which competent users of the language construct complex tokens from these simple units. Performing a series of inductive generalizations, he infers, quite correctly, that all possible tokens of appropriate categories may be combined in accordance with these patterns. He has still experienced only finitely many tokens, but now his knowledge encompasses an infinity of them. We say that he has learned the language's grammar. Nothing untoward is involved in this jump from finite to infinite knowledge. Only two rather familiar epistemic operations seem to be involved: sense experience of a variety of actual tokens and inductive generalization.

With formalized languages, another source of knowledge becomes available. The properties of possible tokens of formal languages are not determined by common usage, as with a natural language; they are stipulated in its grammatical rules. These rules, as we understand them, are stipulations in an abstract language about possible expression tokens, and anything which can be known about the language or languages they define can be deduced from them. Consequently,
the logician who lays down these stipulations and anyone else who understands them can learn a great deal from them simply by deduction.\(^9\)

In practice, however, induction almost always plays a role here too. Making deductions from a set of grammatical rules gives the logician a "feel" for the interpreted language they define. Thus by inductive means he is able to guess what new facts he is likely to be able to deduce and which methods of proof are likely to yield the deductions.

Moreover, someone unfamiliar with such rules—a student, for example—usually needs to work with actual examples of the formulas they define in order to perform the inductions necessary to comprehend their meaning. Thus when a logic teacher presents a set of formation rules to his class for the first time, he generally also gives examples of various expressions and asks students whether these do or do not conform to the rules. Only after seeing examples of this sort and performing inductive generalizations on them do most students understand the rules and thereby become capable of deducing knowledge about other possible tokens of the language.

Thus we have at least three means of acquiring knowledge of nonactual tokens: (i) rendering them actual by construction and then observing them directly, (ii) inductive generalization from experience with actual tokens or with grammatical rules governing them, and (iii)
deduction from grammatical rules.

All three play some role in the acquisition of set-theoretic knowledge. The axiomatizers, the inventors of set-theoretic object languages, get much of their knowledge, as we said earlier, simply by having stipulated the facts they know. Of course, not all of the consequences of their stipulations are immediately evident to them, and so they also learn about the tokens whose possibility they have stipulated by deduction (aided by inductive heuristics) from their axioms.

The student of axiomatic set theory learns about the possible tokens of set theory's object language by being given the axioms (i.e., the incomplete grammatical conventions governing this object language) and being taught how to use them. The axioms are not a complete set of grammatical rules, but they define the language he is to learn about with precision sufficient for his purposes. In particular, the axioms provide no information which would enable him to discern equivalent tokens—but, qua set theorist, he has no use for such information. He comes to know about the possible formulas which the axioms describe in part by using and experiencing these or closely related formulas, in part by induction from this experience, and in part by deduction from the axioms (augmented, again, by inductive heuristics).

It is the need to make the inductive generalizations just mentioned which accounts for the fact that set theory
is learned by solving problems with pencil in hand. The practice the student gains working problems familiarizes him with the breadth and variety of set theory's possible tokens and enables him to perform the inductive generalizations which provide much of his knowledge of merely possible ones.
Since its beginnings in the eighteenth century, modern empiricism has been at pains to explain the nature of mathematical truth. Mill took the heroic position that mathematical statements are just extremely well-confirmed inductive generalizations, but most of his empiricist successors found this view untenable. Mathematics, like logic, they argued, has a peculiar kind of certainty, stronger than any that could be achieved by induction from experience. In the twentieth century there arose a view, first articulated in Wittgenstein's *Tractatus*, that mathematical statements are tautologies, devoid of empirical content. This view reached its fullest expression in the works of the Vienna Circle positivists. The following formulation by A. J. Ayer is typical:

The principles of logic and mathematics are true universally simply because we never allow them to be anything else. And the reason for this is that we cannot abandon them without contradicting ourselves, without sinning against the rules which govern the use of language, and so making our utterances self-stultifying. In other words, the truths of logic are analytic propositions or tautologies ([AYER [1952], p. 77]).
There were a number of variations on this positivistic theme. Mathematics and logic were characterized as "true by definition," "true by convention," "true in all possible situations," "necessarily true," etc. These doctrines were often blended and confused with one another, and little attempt was made to distinguish them.

As evidence for their views, the positivists offered the work of Russell and Whitehead. The authors of *Principia* had derived all known mathematics from a small number of axioms expressed in the language of the higher-order predicate calculus. Most, of all, of these axioms, the positivists held, were analytic or true by convention. Since mathematics followed from them by logical inference and definition alone, it was argued that mathematics, too, inherited their analytic character.

But this "evidence" was problematic. *Principia*'s axioms, the most important of them at any rate, seemed under close scrutiny to be neither true by convention, nor analytic, nor even necessarily true. This was the case, not only for the axioms of reducibility, choice, and infinity—all of which Russell and Whitehead had more or less reluctantly appended to the "logic" of their system in order to obtain various mathematical results—but for that very "logic" itself. The propositional calculus and first-order predicate calculus undoubtedly deserve the title "analytic" if anything does, but the "logic" of Russell and Whitehead goes
far beyond them in power—and, apparently, in ontological commitment. The theorems of first-order logic assert only the existence of one unspecified object (and a slight emendation eliminates even this much "commitment"\(^1\)), but the logic of Principia, classically understood, asserts the existence of an infinite hierarchy of attributes or propositional functions. If there are no such things, this "logic" is apparently false. Therefore, it cannot be analytic (true in virtue of meaning alone), since its truth also depends upon the existence of propositional functions. Furthermore, unless propositional functions necessarily exist, which seems doubtful, the "logic" of Principia is not even necessarily true.

The positivists' case seemed to weaken still further when the unwieldy system of Principia gave way to ZF as the chief contender for the foundations of mathematics. ZF's alleged existence assumptions are blatant—not disguised by the innocent-seeming device of higher-order quantification.\(^2\) It no longer retains even the appearance of analyticity or tautologousness.

In view of our analysis, of course, the positivists' case was not as weak as it appeared. Neither ZF nor the logic of Principia asserts the existence of anything, and each is conventionally true. But the positivists had no such analysis. Faced with the collapse of the arguments for their position, they retreated to extreme and implausible
views about the nature of theoretical existence assertions.

Nowhere is this more apparent than in Carnap's "Empiricism, Semantics, and Ontology" (1950). There Carnap tries to neutralize the apparent ontological commitment of theories about abstract objects by arguing that it is meaningless to speak of what exists and what does not, except from within a specific "linguistic framework." If I say within the framework of number theory, for example, that numbers do not exist, my statement will be meaningful but false; the rules governing the number-theoretic framework dictate that numbers exist. And then, if I protest, "but there really are no such things as numbers," Carnap's reply is that this statement is meaningless, because it is not offered within the number-theoretic framework, but is a broader metaphysical statement which lacks a framework altogether. To introduce a theory of abstract entities, Carnap holds, is not to commit oneself to the metaphysical reality of such entities, but simply to accept a new way of speaking:

...the introduction of the new ways of speaking does not need any theoretical justification because it does not imply any assertion of reality. We may still speak (and have done so) of "the acceptance of new entities" since this form of speech is customary; but one must keep in mind that this phrase does not mean for us anything more than acceptance of the new framework, i.e., of the new linguistic forms. Above all, it must not be interpreted as referring to an assumption,
belief, or assertion of the "reality of the entities." There is no such assertion. An alleged statement of the reality of a system of entities is a pseudo-statement without cognitive content (p. 241).

Carnap's view might have fared better had he held that the apparently abstract entities introduced by new linguistic frameworks like set theory or number theory are not asserted to exist, but only to be possible. Instead, however, he clung to the idea that the existential quantifications of such theories were in fact existence assertions—but existence assertions which were somehow not to be taken too seriously, existence assertions with no ontological significance.

The ground was now prepared for a devastating assault on the positivistic philosophy of mathematics. And the assault came—on a number of fronts—in a series of papers by Quine. An immediate target was the positivists' ontological frivolousness. Wielding his famous criterion of ontological commitment, Quine charged that to fail to take one's existence assertions seriously was to fail to make oneself understood:

...consider the man who professes to repudiate universals but still uses without scruple any and all of the discursive apparatus which the most unrestrained of platonists might allow himself. He may, if we train our criterion of ontological commitment upon him, protest that the unwelcome commitments which we impute to him depend on unintended interpretations of his
statements. Legalistically his position is unassailable, as long as he is content to deprive us of a translation without which we cannot hope to understand what he is driving at (QUINE [1961], p. 105).

There was little argument for the criterion itself, but the idea was simple. We commit ourselves to the existence of something when we say that it exists, and we say it exists whenever we quantify over it. For those who sought exemption from the criterion, Quine offered a dilemma:

To insist on the correctness of the criterion...is, indeed, merely to say that no distinction is being drawn between the 'there are' of 'there are universals', 'there are unicorns', 'there are hippopotami', and the 'there are' of '(\exists x)', 'there are entities such that'. To contest the criterion, as applied to the familiar quantificational form of discourse, is simply to say either that the familiar quantificational notation is being re-used in some new sense (in which case we need not concern ourselves) or else that the familiar 'there are' of 'there are universals' et al. is being re-used in some new sense (in which case again we need not concern ourselves) (Ibid.).

Now this dilemma was effective against Carnap and others, who never denied that their quantifiers meant "there are." (As for Carnap's claims of meaninglessness, Quine offered separate arguments, which need not be detailed here.) But it would have had far less impact if the positivists had replied that they were indeed using the quantifiers in an untraditional sense—a modal sense. Quine, of course, would still have insisted that he could not understand what was
being said, since it is untranslatable into his "canonical idiom." But we have already adduced reasons for believing contra Quine that modal quantification can be readily understood—at least in some contexts—so that one could reply that there are intelligible statements which cannot be made within the confines of a language whose quantifiers are understood only classically.

Problems of ontological commitment, however, were just one thrust of Quine's many-pronged attack. A more direct assault on the idea that logic and mathematics are true by definition or convention came in his "Truth by Convention" (1936b). There Quine questions the very intelligibility of the idea of conventional truth. If we allow ourselves appropriate conventions, he argues, even "empirical" sciences can be made true by convention. He offers a simple recipe for doing so:

In each case we merely set up a conjunction of postulates for that branch [of science] as true by fiat, as a conventional circumscript of the meanings of the constituent primitives, and all the theorems of the branch thereby become true by convention...
(p. 92).

The fact that any theory can be made conventionally true in this way now poses a three-pronged dilemma:

If in describing logic and mathematics as true by convention what is meant is that the primitives can be conventionally circumscribed in such a fashion as to generate all and only
the accepted truths of logic and mathematics, the characterization is empty; our last considerations show that the same might be said of any other body of doctrine as well. If on the other hand it is meant merely that the speaker adopts such conventions for those fields but not for others, the characterization is uninteresting; while if it is meant that it is a general practice to adopt such conventions explicitly for those fields but not for others, the first part of the characterization is false (p. 95).

Quine's conclusion is that no one has given an acceptable elucidation of the notion of truth by convention, and he suggests darkly that the idea itself lacks sense:

...as to the larger thesis that mathematics and logic proceed wholly from linguistic conventions, only further clarification can assure us that this asserts anything at all (p. 99).

Quine has another problem for the more specific thesis that logic is true by convention. There is a regress in setting up the conventions; we cannot stipulate what is true by convention in logic without resorting to the very logical laws whose conventionality we are trying to establish. Thus if we make the laws of propositional logic conventionally true by stipulating that some appropriate axiom set and all its consequences are true, we cannot say what those consequences are unless we assume the truth of the very logical principles we are trying to stipulate. Now there is a way out of this regress, and Quine discusses it, but he thinks that it deprives the notion of truth by
It may be held that we can adopt conventions through behavior, without first announcing them in words; and that we can return and formulate our conventions verbally afterward, if we choose, when a full language is at our disposal. It may be held that the verbal formulation of conventions is no more a prerequisite to the adoption of the conventions than the writing of a grammar is a prerequisite of speech; that explicit exposition of conventions is merely one of many important uses of a completed language. So conceived, the conventions no longer involve us in a vicious regress. Inference from general conventions is no longer demanded initially, but remains to the subsequent sophisticated stage where we frame general statements of the conventions and show how various specific conventional truths, used all along, fit into the general conventions as thus formulated.

It must be conceded that this account accords well with what we actually do. We discourse without first phrasing the conventions; afterwards, in writings such as this, we formulate them to fit our behavior. On the other hand it is not clear wherein an adoption of the conventions, antecedently to their formulation consists; such behavior is difficult to distinguish from that in which conventions are disregarded. When we first agree to understand 'Cambridge' as referring to Cambridge in England, failing a suffix to the contrary, and then discourse accordingly, the role of linguistic convention is intelligible; but when a convention is incapable of being communicated until after its adoption, its role is not so clear. In dropping the attributes of deliberateness and explicitness from the notion of linguistic convention we risk depriving the latter of any explanatory force and reducing it to an idle label. We may wonder what one adds to the bare statement that the truths of logic and mathematics are a priori, or to the still barer behavioristic statement that they are firmly accepted, when he characterizes them as true by convention in such a sense (pp. 98-9).
Quine's offensive did not end with the attack on convention. In "Two Dogmas of Empiricism," he directed his fire against two bulwarks of positivism—the analytic-synthetic distinction and the idea that each meaningful statement is reducible to a statement whose terms refer to immediate experiences. The latter "dogma" does not concern us, but the former is central to the positivistic philosophy of mathematics. The positivists had held that a neat analytic-synthetic distinction could be drawn, with mathematics and logic on one side and the empirical sciences on the other. But this view was called into question, as we saw, with the realization that Principia's axioms—and, later, the axioms of ZF—seemed to be substantive assumptions. This blurring of the distinction paved the way for Quine's denial of it.

The strategy of "Two Dogmas" is much the same as that of "Truth by Convention." Quine argues that the proponents of the analytic-synthetic distinction, like the proponents of truth by convention, have failed to give their idea any intelligible sense. He then proceeds by elimination, examining a number of different accounts of analyticity and rejecting each in turn. His conclusion is that there is no distinction to be made; the difference between the analytic and the synthetic (if these terms may still be used) is one of degree, not of kind. Statements more readily subject to denial in the face of recalcitrant data are more nearly
synthetic; statements central to our conceptual scheme, and hence more firmly accepted, are more nearly analytic. But no statement is totally immune to rejection—or purely analytic.

This means for Quine that even mathematics and logic are scientific theories, no different in kind from the empirical sciences. Set theory, as well, becomes an empirical investigation, guided by the same theoretical principles as any investigation of the world:

The issue over there being classes seems more of a question of convenient conceptual scheme; the issue over there being centaurs, or brick houses on Elm Street, seems more of a question of fact. But I have been urging that this difference is only one of degree, and that it turns upon our vaguely pragmatic inclination to adjust one strand of the fabric of science rather than another in accommodating some particular recalcitrant experience. Conservatism figures in such choices, and so does the quest for simplicity (QUINE [1961], p. 46).

Sets, like muons, are objects whose existence is revealed to us as we organize and systematize the data of our experience; and set theory is, just as it appears, a substantive theory of the world.

Now this, of course, is precisely the view we found so problematic in Chapter 1. Muons are spatially and causally related to other objects in the world, and we can readily catalogue the effects by which they manifest themselves and by which we know about them. But sets apparently
have no such tie to the physical world, and so their epistemology and metaphysics remains blanketed in obscurity. The analysis we have developed since rejecting this view has much in common with the positivistic philosophy of mathematics. Set theory is true, on this analysis, not in virtue of any feature of the world, but because its axiomatizers adopted certain conventions (axioms) governing the structure of language hierarchies and the behavior of the satisfaction predicate. It is therefore true by convention. Since conventions are circumscriptions of meaning, it is also true in virtue of meaning, i.e., analytic; and since conventions do not change from world to world in evaluating the truth of statements, it is also true in all possible worlds, i.e., necessary.

Finally, it is, as the positivists urged, devoid of factual content and hence unfalsifiable; for, once conventions are adopted, if consistent, they remain true regardless of any changes in the world. Of course, they may later cease to be observed or thought about, but to abandon them is not to falsify them; it is simply to trade one form of language for another—or for none at all. We may abandon twentieth century English, but no conceivable event could make the statement "It is sunny" is a possible sentence of twentieth century English' false. Similarly, we may abandon ZF, but the statement \( \sim \emptyset \in \emptyset \) of ZF remains true for all time (assuming ZF is consistent), since it is the
consequence of the conventions which make ZF what it is, in the same way that the statement above is a consequence of the conventions of twentieth century English.

Yet our analysis diverges in some respects from the positivistic view. The early positivists claimed that mathematics is "true by definition." This phrase was used loosely, but often what it meant was that mathematical truths were definitional transcriptions of theorems of logic. Clearly set theory as we see it is not true by definition in this sense, since its postulates (e.g., the axiom of choice) go far beyond those definitionally derivable even from higher-order logic.

Nevertheless, our analysis is committed to many of the "dogmas" which Quine rejects. Chief among these is the doctrine of conventional truth. Quine's main objection to this doctrine is his tripatite dilemma, which runs as follows:

(1) If what the doctrine means is that logic and mathematics can be made conventionally true, then it is empty; any other theory can as well.

(2) If what it means is that the speaker himself holds logic and mathematics to be conventionally true, then it is uninteresting.

(3) If what it means is that "it is a general practice to adopt such conventions explicitly for those fields but not for others, the first part of the characterization is false."
The first two points are perfectly correct. Any theory can be made true by convention or can be held to be conventionally true by some individual. But these facts are of little interest. The question of importance with respect to a live theory is not whether it can be made conventionally true, nor whether someone holds it to be conventionally true, but whether it is conventionally true as it is used in actual practice.

The third point is partly right, but mostly wrong. Logic, of course, was not adopted by explicit stipulation. (Quine's regress argument shows this to be impossible.) But set theory, as we understand it, was; formulation of its axioms amounted to explicit stipulation of grammatical rules. Moreover, the fact that logic was not adopted by explicit stipulation does not mean that the conventions by which it is true are not more or less explicit today; they are, as we shall see shortly. However, none of these objections to (3) gets to the heart of the matter; for, like (1) and (2), (3) does not capture what we mean by truth by convention. There is a fourth alternative, the correct one, which Quine has ignored altogether.

A major selling point of Quine's argument is that any theory can be made conventionally true, and he gives us an algorithm for doing the job. But it is important to see that no theory is interpreted in practice by the artificial "grammar" which Quine proposes. Truth is assigned to
formulas by recursion on complexity as a function of the semantic interpretation of primitives, not on the basis of deducibility from a set of axioms. One could, no doubt, axiomatize zoology and stipulate that the axioms and all their consequences are true, but this is not what we do in fact. We employ no grammatical rule which implies, for example, that the sentence 'Blue whales eat plankton' is true just in case this sentence follows from axioms X. The interpretation rules governing all the languages we actually use are, instead, Tarski-style truth definitions—perhaps with some additional meaning postulates, which we shall discuss momentarily. Such rules imply, for example, that 'Blue whales eat plankton' is true if and only if blue whales eat plankton, and these are the appropriate truth conditions for that sentence. Zoology is not conventionally true in actual practice, because the conventions governing it in actual practice do not entail its truth. To deduce the truth of its theorems, we would need additional postulates which imply, among other things, that blue whales eat plankton.

Zoology, of course, does have some conventional truths. Its truth-functionally tautologous statements are the most obvious examples, and these shade off into truths about which there is more controversy. Such, for example, is the statement 'All cats are animals'. (See PUTNAM [1962a].) This statement seems to function as a meaning
postulate, and its logical consequences may be thought of as inheriting truth-value in something like the way Quine described. We still, of course, interpret these consequences truth-functionally after the fashion of Tarski, so that our grammatical rules imply, for example,

\[(4) \ 'All cats are animals or blue whales eat plankton' is true iff all cats are animals or blue whales eat plankton.\]

But we also seem to have a separate grammatical rule or meaning postulate which stipulates

\[(5) \ All cats are animals,\]

so that the statement

\[(6) \ 'All cats are animals or blue whales eat plankton' is true\]

is deducible from grammatical rules alone.\(^9\)

This example, as we noted, is controversial; and it may be, as Putnam and others have argued, that there is some indeterminacy about just what the conventions of zoology are. But, assuming that at least some sense can be made of the notion of the grammatical rules governing a language in practice (and the entire science of theoretical linguistics assumes this), then we can say what the interesting notion of truth by convention comes to fairly clearly. A theory is true by convention in practice only if the grammatical
rules governing it in practice imply that each of its theorems are true.

While following from appropriate grammatical conventions is a necessary condition of conventional truth in practice, it is not sufficient. For if the grammatical governing a language are inconsistent (as was the case with Cantor's satisfaction theory), then falsehoods will follow from them. Accordingly, a theory is true by convention in practice if and only if

(i) the grammatical rules governing it in practice imply that each of its theorems are true, and

(ii) these rules are consistent.

This is the fourth alternative, which Quine fails to mention. Notice that it does not require that the conventions be explicit, as (3) does.

Even allowing for some indeterminacy in the grammatical rules of the languages of natural science, it is clear that no natural science is conventionally true in practice on this definition. No case can be made for the claim that

(7) 'Blue whales eat plankton' is true

follows from any grammatical rules governing the current language of zoology. The concept of truth by convention is not, therefore, completely without meaning as Quine suggests.
Quine, however, would not be happy with what we have said here. For reasons related to his doctrine of indeterminacy of translation, he does not think that the notion of a grammatical rule employed in practice is even intelligible. To resolve this matter—especially for the natural languages in which the sciences are usually formulated—would require a lengthy discussion beyond the scope of this work. But our concern is not with theories of natural science, nor with theories couched in ordinary language; it is with axiomatic set theory expressed in the formal language $\mathcal{L}^S$. For such a simple and artificial language we may hope to make the notion of a grammatical rule employed in practice fairly clear, even by Quinean standards.

Semantical rules, as we suggested before, are of two types—Tarski-style truth definitions and meaning postulates. (Explicit stipulative definitions are a rather uncontroversial example of the latter.) Now it is uncontestable, it seems to me, that Tarski-style truth definitions are the semantical rules governing the predicate calculus in practice and hence also the rules governing the logical operators of $\mathcal{L}^S$. Tarski-style semantics is nowadays universally accepted by logicians as the standard semantics for classical first-order logic. Show a logician or mathematician a set of Tarski-style truth definitions for his language and ask him if these are the rules of interpretation which he is using, and you will immediately get a
positive reply. These conventions are not merely followed; they are printed in textbooks and explicitly accepted by the logical and mathematical communities at large.

Now the truth of each of the theorems of standard first-order logic follows from Tarski-style truth definitions. (This is what we show, in effect, when we prove that the theorems of first-order logic are valid, i.e., true in all models.) Therefore, first-order logic is true by convention.

In retrospect we can understand why Quine, writing "Truth by Convention" in 1936, would have doubted this. For one thing, it was just in that year that the grammatical conventions governing formal first-order languages were first made fully explicit, in TARSKI (1936). These conventions, as we shall argue shortly, had been implicit in the use of these languages all along, but they were not until then explicitly and formally stated.

Quine was also impressed, perhaps overly so, with his regress argument. Conventions for logic, he reasoned, cannot be stated without use of logic, so that logic could not first have become true by explicit stipulation, since any such stipulation would presuppose the truth of logic. But notice that Quine's premises only support the conclusion that logic could not first have become true by explicit stipulation; as Quine himself admits, they do not imply that the conventions governing logical languages cannot be
adopted explicitly at some later stage. And that, as we have indicated, is precisely what has happened. Since Quine wrote "Truth by Convention," Tarski-style truth definitions have become universal fare among logicians and appear even in elementary texts. Thus Quine's claim in (3) that logic is not true by explicit convention, though true in 1936, is no longer true today.

Quine wonders what sense can be given to the notion of truth by convention beyond "the bare statement that the truths of logic and mathematics are a priori, or...the still barer behavioristic statement that they are firmly accepted." With respect to logic, at least, we now have an adequate answer. A theory is true by convention if its truth follows from the conventions governing it in practice (provided that these conventions are consistent); and there are rigorous, well-defined, consistent, and explicitly accepted conventions which imply the truth of the theorems of logic.

The fact that logical languages are now governed by explicit grammatical rules is good evidence that they were implicitly governed by these same conventions even before the conventions themselves were formulated. Tarski's formulation of them did not change, but merely made explicit, the way we ordinarily assign truth in classical formal languages. Logicians recognized immediately that Tarski's characterization was accurate. Thus not only is classical
logic now true by convention, it apparently always has been—at least implicitly.

It remains to be shown that set theory—and, in particular, ZF as expressed in $L^S$—is conventionally true. To see this, we must once again examine the role played by ZF's axioms.

We argued in Chapter 6 that the axiomatization of set theory was, in effect, the reformulation of that part of the language of the theory of satisfaction which deals with unary formulas. Our evidence was based on the explanatory power and metaphysical level-headedness of this hypothesis. The advantages of the hypothesis are many. It avoids the assumption that the axiomatizers were engaged in describing a mystical realm of platonic things. It accounts for the fact that set theorists make no use of observational data and the fact that their work consists in manipulation of formulas. It explains the similarity between the "logical" and "semantic" paradoxes and goes a long way toward accounting for set theory's fundamental principles, instead of regarding them as brute facts. It allows us to hold that the axiomatizers justifiably took the "existence" of sets for granted and that their enterprise was not in violation of Occam's razor. And, most importantly, it renders the set-theorist's knowledge explicable. On the basis of this systematic evidence, we felt justified in concluding that the axiomatization of set theory amounted to reconstruction.
of the language of the theory of satisfaction.

It is important in understanding this conclusion to distinguish the language in which the axioms themselves are formulated ($L^S$) from the language (really a hierarchy of languages) of the satisfaction theory. $L^S$ is the metalanguage for the hierarchy, and the axioms state truths about its possible unary formulas. For example, the union axiom "says" that for any unary formula of the hierarchy, there is possible a second unary formula satisfied by just those unary formulas which satisfy the unary formulas satisfying the first. Thus the axioms codify syntactic and semantic truths about the possible formulas of the hierarchy in the same way that formulation and interpretation rules do, except that they are less complete than the latter. Thus the union axiom, though it tells us that for each unary formula, it is permissible to construct a second (syntactically) distinct formula having certain semantic properties relative to the first, does not specify the syntactic structure of either formula.

We concluded in Chapter 6 that the role played by the axioms was to define (albeit in this vague and incomplete form) the structure of the new semantic hierarchy which was to replace the inconsistent semantics implicit in Cantor. Their practical function was to stipulate what is and is not possible (i.e., grammatical) within this new hierarchy. And this is precisely analogous to the function of formation and
interpretation rules in setting up a formal language. Such rules are not descriptions of some antecedently given reality. Rather, like meaning postulates, they set limits to our use of language. Just as the meaning postulate 'Triangles have three sides' expresses a convention about the sorts of things we may regard as triangles, and just as the grammatical rules of a formal language express conventions about what are to be regarded as its possible formula tokens, so the axioms of ZF express conventions about what are to be regarded as possible unary formulas of the new semantic hierarchy (or, if you like, about what are to be regarded as "ZF-sets"). Their truth, like that of all rules of grammar, is conventional.

And being conventions, the axioms are true by convention in our sense. For the grammatical rules governing $L^S$ include not only Tarski-style truth definitions, which imply

\[(8) \ A \text{ is true iff } tr(A)\]

for all formulas $A$ of $L^S$, but also all the axioms themselves (or, rather, their translations into English). Since the axioms are statements of conventions of $L^S$, so are their translations into its metalanguage, so that for each axiom $B$, \[\text{"}tr(B)\text{"}\] is a conventional truth of English. Thus for each axiom $B$, the grammatical rules governing $L^S$ imply

\[(9) \ B \text{ is true,}\]
from which it follows from our definition of truth by con-
vention that ZF (if consistent) is conventionally true. It
also follows, from considerations mentioned previously, that
ZF is analytic, necessary, and true in all possible worlds.

But why, the reader may ask, is set theory special
in this regard? In what way do its axioms function differ­
ently from those of, say, quantum mechanics? The answer is
already implicit in the preceding discussion. The axioms
of set theory function in practice as stipulations of what
is grammatical (possible) in certain languages, while the
axioms of quantum mechanics seem to function, at least in
part, to describe what is actual. 10 Natural science is
apparently a tangle of empirical truth and convention; and
it may be, as Quine and Putnam have argued, that these two
strands are interwoven inextricably. But set theory, as our
analysis views it, has only one of the strands. It is pure
grammar, pure convention. Like any system of grammatical
rules, it may prove unwieldy, fall into disuse, even be
forgotten; but it cannot be falsified, except by derivation
of a contradiction. Just as consistent formation and inter-
pretation rules for a formal language are not subject to
disconfirmation, so set theory (if consistent) would remain
true even in oblivion.

If set theory is conventionally true, then those
statements whose negations are theorems are conventionally
false. But what of those statements A such that neither A
nor $\sim A$ follows from the axioms? As we said earlier, incomplete rules define, not a single language (or language hierarchy), but an entire range of languages. Statements independent of the axioms are true for some members of this range and false for others. Now if the founders of set theory had had some one member of the range defined by their axioms in mind (i.e., if they had intended to complete the axioms in a certain way), then we could assign truth-values to these sentences in accordance with their intentions. But they did not. It is obvious, of course, that the rules they formulated are incomplete with respect to such matters as details of syntax. $L^S$ contains no syntactic predicates, and so rules directly governing these matters cannot even be formulated. Indeed, we can profitably regard the axiomatizers as thinking in an abstract language akin to $L^S$, a language from whose point of view syntax is largely irrelevant and in which unary formulas "appear" as sets. Now it might be thought that, viewed through the spectacles of this more abstract language, the structure they were thinking of appeared determinate (i.e., that they had in mind some determinate way of completing the axioms within $L^S$). But even this is not the case. What they envisioned through their abstract spectacles is what would nowadays be called the standard model. The structure of the standard model has never been—and cannot be—completely characterized by finite means. Since it seems safe to assume that their
intentions were finite (i.e., that they corresponded to
some finite set of statements), we must conclude that even
their intentions were incomplete. They had in mind no one
determinate way of completing the axioms.

Questions about the truth-value of statements inde­
dendent of the axioms are rather like the question of whe­
ther Oedipus was born on the Fourth of July. The Oedipus
myth, which is the only arbiter of truth in this case,
fails (I assume) to supply an answer. Likewise, the axioms
of ZF, which are the only arbiters of truth about the
semantic hierarchy they define (unless we count the axioma­
tizers' intentions, and even these lack completeness), fail,
for example, to tell us whether there is a measurable
cardinal.

The comparison is illuminating in several respects.
Just as the axioms in their incompleteness succeed in defi­
ning only a range of possible hierarchies, so the Oedipus
myth, similarly incomplete, defines, not a single life-story
for Oedipus, but many; in some of these stories Oedipus was
born on the Fourth of July, in some he was born on the
fifth, and so on. Further, the Oedipus myth is not (or not
be) a fixed set of statements. In it heyday, no doubt, it
was a growing and evolving tale. More detail was added with
time, so that the range of possible life-stories defined
became smaller. Set theory evolves in much the same way.
Over time, more axioms gain wide acceptance—as the axiom of
choice, and perhaps also the continuum hypothesis, have today. With each new accretion, the range of possible hierarchies defined narrows. Therefore, though our present axioms leave the question of whether there is a measurable cardinal unresolved, this question may well be resolved in the future—as the plot of our story, so to speak, thickens.

It is best, then, to regard statements independent of the axioms as without truth-value, at least for the moment. There is no fact of the matter about their truth.

Admission of truth-value gaps does not entail abandonment of classical logic. We discussed two methods of dealing with truth-value gaps in Chapter 5. Of these, the approach of KLEENE (1962) requires changes in classical logic, but that of VAN FRAASSEN (1966) does not. Van Fraassen's approach amounts to arbitrarily closing truth-value gaps—or, in terms of our metaphor, to arbitrarily completing the story—for purposes of reasoning. Thus, for example, the sentence of \( L^S \) expressing "either there is or is not a measurable cardinal" is a logical truth, even though the truth-value of each disjunct is currently indeterminate. And this seems rather natural; for on every way of completely elaborating our story (i.e., for every possible hierarchy in the range which the axioms define) this sentence is true.

It should be clear by now that on our analysis truth in set theory amounts to provability. The true sentences of
... are those deducible (by classical logic, if we follow van Fraassen) from the axioms, and the false ones are those whose negations are deducible. The others lack truth-value.\textsuperscript{12}

This reaffirms the connection between possibility and provability discussed in Chapter 2. We said there that when a set $R$ of incomplete grammatical rules (e.g., the axioms of ZF) defines a range of possible languages or language hierarchies, tokens satisfying a condition $A(x)$ are possible for each member of this range just in case $R \vdash \exists x A(x)$ and not possible for each member just in case $R \vdash \neg \exists x A(x)$. If $\exists x A(x)$ is independent of the axioms, then tokens satisfying $A(x)$ are possible in some members of the range but not in others. Since there is no fact of the matter about which member of the range is the intended one, however, the only "existentially" quantified sentences actually given truth-value by our conventions are those which are not independent of the axioms. Thus an assertion of possibility in set theory (i.e., an "existentially" quantified statement) is true just in case it is provable and false just in case its negation is. If it is independent of the axioms, then its truth-value is indeterminate.
By now it is evident that the attenuation of the notion of a possible formula token required by our analysis has reached an advanced state. The possible tokens over which set theory quantifies belong to a language or language hierarchy which is defined only loosely and is not even governed by recursive formation or interpretation rules. The nature of these possible formula tokens is therefore very imprecisely determined, and their syntactic structure is not specified at all. Since their syntax is so indefinite, we might think of these "formulas" as possible utterances, possible notches in wood, possible mental concepts—indeed, possible objects of almost any sort, so long as they are capable of being assigned the sorts of satisfaction conditions which the axioms stipulate.

It might seem, then, that our analysis collapses into the view of Putnam in "Mathematics without Foundations" (1967):

...there is not, from a mathematical point of view, any significant difference between the assertion that there exists a set of integers satisfying an arithmetical condition
and the assertion that it is possible to select integers so as to satisfy the condition. Sets, if you will forgive me for parodying John Stuart Mill, are permanent possibilities of selection (p. 12).

Indeed, we have now stretched things so far that possible formula tokens for us seem to be little more than Putnam's permanent possibilities of selection—hardly identifiable as formulas at all.

Yet we have not gone quite that far. Set theory is a body of knowledge obtained by inductive generalization from observation of actual formula tokens and by deduction from axioms which function in many ways like grammatical rules. Thus to think of sets merely as permanent possibilities of selection is to miss their crucial link with actual formula tokens, the objects which we use in practice to count, categorize, calculate, and make "selections."

Formally, we can regard sets as objects of any sort we please; but the "sets" that we actually work with (i.e., unary formula tokens) are all inscriptions or utterances of a fairly conventional sort. Therefore, it is most reasonable to regard the nonactual objects over which set theory quantifies as like them in kind. In a sense, sets are permanent possibilities of selection; but the gist of our argument is that permanent possibilities of selection are most profitably understood as possible formula tokens, or possible objects not radically unlike them.
Our concern thus far has been with unary formulas alone, but we suggested in Chapter 1 that set theory could be understood as part of a larger semantic theory governing possible tokens with any number of free variables. Such a theory would need something like the variably polyadic predicate $T$ described there, and could be constructed either in the form of a language hierarchy or in the manner of Kripke's theory. The $\epsilon$ of set theory could be introduced by the definition

\[(1) \ x \in y \text{ for } yT(x),\]

and the predicates $=$ and $\equiv$ could be introduced by DEF 3.1 and DEF 3.2 (or their type-theoretic analogues, if the language contained variables of various types). Extensionality for unary formulas would be expressible by the usual axiom or axiom schemata. We could also add a predicate expressing extensional equivalence for formulas with $n$ variables:

\[(2) \ x \equiv y \text{ for } \forall z_1 \ldots \forall z_n (xT(z_1, \ldots, z_n) \leftrightarrow yT(z_1, \ldots, z_n)).\]

If, in addition, we incorporated a way of distinguishing formulas with different numbers of free variables, this predicate could be used to express a general schema of extensionality. Suppose, for example, we have the unary predicates $V^n$, where $V^nx$ means 'x has n free variables'.
Then we can write the schema as

\[ (3) \forall x \forall y((\forall^{n}x \land \forall^{n}y) \rightarrow (x = y \leftrightarrow x \subseteq y)). \]

On analysis, the schema asserts that formulas \( A \) and \( B \) with \( n \) free variables are indiscernible just in case for any objects \( a_1, \ldots, a_n \), \( A \) is true (of \( a_1, \ldots, a_n \)) if and only if \( B \) is.\(^2\) If we construct and interpret our language so that this schema is true, then equivalent formulas in \( n \) free variables become indiscernible for each \( n \). Thus, just as extensionally equivalent unary formulas "appear" under this sort of abstraction as sets, so do extensionally equivalent formulas in many free variables "appear" as relations-in-extension. Materially equivalent sentences also become indiscernible, so that there appear to be only two objects satisfying the formula \( \forall^0 x \). Intuitively, these are the things Frege reified under the titles "the True" and "the False." Thus expansion of our semantic theory permits analysis of other "abstract entities" besides sets.

Treatment of relations-in-extension as formulas in more than one free variable, though in a sense less parsimonious than the standard set-theoretic treatment, is in some ways more natural. The various definitions of ordered pairs due to Kuratowski, Weiner, Quine (1945), and others, all seem rather artificial and ad hoc; and the treatment of functions and relations based upon these definitions, while formally impeccable, strikes many people as only distantly
related to concepts of relations or functions employed in ordinary or philosophical discourse. Treatment of relations as functions in abstraction is much closer in form to the well-motivated treatment of Frege and Russell, and is perhaps more satisfying philosophically than the standard mathematical analysis. Relations are simply actual or possible bits of language which we use or could use to draw connections among objects.

The fact that sets, relations-in-extension, and truth-values (i.e., the True and the False) all emerge on our analysis as possible formulas spoken of in an abstract idiom points to a rather novel reinterpretation of some semantical doctrines stemming from Frege. Frege thought that objects of these sorts were the denotations respectively of unary predicates, relational predicates, and sentences. The notion of denotation here is understood by analogy to the denotation of singular terms; denotations are nonlinguistic objects which the linguistic ones name. We might call this a "word-object" view of the semantics of predicates and sentences.

Our analysis gives a very different picture. Suppose, for example, we interpret a unary predicate $F$ of some formal language by assigning it the "denotation" $\{x \mid x \text{ is a fox}\}$, i.e., we give it the truth conditions

\[(4) \forall y (Fy \iff y \in \{x \mid x \text{ is a fox}\}).\]
What we have done in effect is simply to stipulate the extensional equivalence of $F$ to the predicate 'x is a fox' and its equivalents—the objects denoted indiscernibly by the term '{x | x is a fox}'. Assignment of abstract denotations like sets, relations, and truth-values to predicates and sentences therefore amounts simply to equating one predicate or sentence with others. It establishes links only between words and words—not, as the Fregean picture suggests, between words and abstract aspects of the world.

Our treatment suggests a generalization to intensional entities—properties, relations-in-intension, and propositions—as well. These can be regarded respectively as unary formulas, formulas in many free variables, and sentences, all seen from the point of view of a language slightly less abstract than the one we have been discussing—a language in which all and only synonymous or logically equivalent expressions are indiscernible. And, again, assignments of intensions to bits of language should be understood, not as establishing connections between language and abstract meanings, but simply as establishing synonymy or logical equivalence between one item of language and others. With respect to intensional objects, as well as extensional ones, our analysis suggests replacement of Frege's word-object picture with a word-word one. In all these cases the main advantage of our analysis is that abstracta are eliminated in favor of possible concreta, some of which are actual.
This foothold on actuality permits inductive generalization which helps to explain knowledge of the nonactual ones.

These brief suggestions do not, of course, constitute an analysis of the objects of Fregean semantics; working out the details is a project far beyond the scope of this study. But one measure of the worth of an idea is generalizibility, and these considerations seem to indicate that our analysis of set theory does possess that virtue.

Mention of various degrees of abstraction brings into sharp focus a question we have so far considered only indirectly. If all these abstract entities are to be regarded as formulas seen through the spectacles of languages of various degrees of abstraction, then how is it that we seem to talk about them all from within one language—our mother tongue, English? The germ of an answer was given in Chapter 2. Though we speak of all these things in English, we do not use all of English to speak about them. Application of certain predicates in abstract discourse creates category mistakes, oddity and confusion. Meaningful talk about these objects takes place only from within fairly well-defined fragments of English, not from within the language as a whole.

But what of cases where we are operating on two or more levels of abstraction at once—such as when we speak of sets of formula types or functions from formula types to physical things? Here again, the principle is the same.
Our language is fragmented, in effect, into two or more sublanguages, each with its own identity predicate.

Formally, this sort of fragmentation can be indicated by the use of different styles of variables—one for sets, one for types, and one for physical objects—and allowing only the appropriate predicates to combine with variables of a given type. Ill-formed combinations correspond to category mistakes. Alternatively, we can retain one style of variables and make the semantics of predicates sensitive to context in such a way that, for example, the predicate 'contains three symbols' is true of \( x \in x \) considered as a type but false (or perhaps undefined) of it considered as a set. These alternatives correspond to the two ways of avoiding metaphysical problems in talk about abstracta or possibilia, which were discussed in Chapter 2.

One area in which our analysis could provide some badly-needed metaphysical house-cleaning is that of modal semantics. For the use of set theory in conjunction with the apparatus of possible worlds generates some untidy puzzles.

One such puzzle concerns sets of possible worlds. If we take the worlds apparatus seriously, then any possible objects exists in some possible world. Now modal semanticians often make use of sets of worlds, sets of sets of worlds, and so on. Each such set must be in some possible
world if it is a possible object at all, but this "nesting" of sets of worlds in worlds seems strange and counterintuitive. And consequences worse than ruffled intuitions follow if our set theory is classical and if, as is often done in modal semantics, we treat worlds as sets and the relation of "being in" a world as membership. For then the existence of the set of all worlds—or, indeed, of any set of worlds which "is in" one of the member worlds—contradicts the axiom of foundation. One could, of course, simply admit the conclusion that no such sets exist, but the price of this admission could be considerable. For the modal semanticist is accustomed to freely invoking such objects as the set of all worlds, and restricting this freedom might render his work significantly more cumbersome—if not impossible.

A second difficulty arises from the clash of the intuitions that (i) set theory is necessarily true, and (ii) a world devoid of all objects is possible. For (i) together with a platonistic reading of set theory implies that sets exist in every possible world. But (ii) denies it. The platonist might abandon (i), but in so doing he risks making the use of set theory in modal semantics unintelligible. For if set theory is true only in a limited class of worlds, then we cannot reason set-theoretically about the "contents" of the others (since set theory does not apply to them), and there seems to be
no reason to believe that set theory cannot legitimately be brought to bear on the worlds apparatus generally. The other option is to deny (ii), but it is hard to imagine independent grounds for doing so, and so the platonist opens himself here to the charge of ad hoc-ness.

Our analysis makes sense of these quandaries. There is nothing especially puzzling about sets of worlds existing in worlds, since what this means is that formulas satisfiable by worlds are possible in those or other worlds. The set of all worlds, for example, exists in the actual world; for what this means is only that formulas extensionally equivalent to 'x is a possible world' are possible in the actual world. There is no temptation to think of sets of worlds as impossible or quasi-possible entities existing somehow outside of all worlds.

The analysis makes it clear, too, that possible worlds—and particularly the actual world—are not sets, since worlds are not formula tokens seen from the point of view of an abstract language. Thus the threatened conflict between classical set theory and the nesting of sets of worlds in worlds does not arise.

That worlds are not sets has a certain intuitive appeal even prior to our analysis. Speaking of Quinean ersatz worlds—complex set-theoretic constructions from actual objects—David Lewis writes:
I cannot believe (though I do not know why not) that our own world is a purely mathematical entity. Since I do not believe that other worlds are different in kind from ours, I do not believe that they are either ([1973], p. 90).

And at first blush this seems right. When we imagine possible worlds—nonempty ones, at least—we do not think of set-theoretic constructions, but rather of collections of possible objects, just as the actual world is a collection of actual objects.

The second problem—the conflict of set theory’s necessity with the possibility of an empty world—simply dissolves on our view. There is a conflict only if set theory makes assertions of actual existence, and we hold that it does not. The possible tokens over which it quantifies still, of course, reside in some possible world or worlds, but they need not reside in every one. Thus it is perfectly conceivable for set theory to be true in all worlds, even though some worlds contain no "sets" (i.e., formula tokens) or are even totally empty.

The past few decades have seen an increasing tendency among metaphysicians to identify a variety of entities with sets. Just now we mentioned construal of worlds as sets, and in Chapter 1 we decried conflation of physical collections with sets of their parts. Quine analyses many objects—including expression types—as sets of various sorts. But the most exotic examples of this metaphysical
genre are to be found in MONTAGUE (1960), CRESSWELL (1973), and related works, where a panoply of intensional entities receive set-theoretic treatment. Cresswell not only holds that worlds, properties, propositions, and the like, are sets; he also explicitly claims that physical objects are functions from worlds to worlds. 6 (He allows worlds to be subclasses of worlds and counts the manifestation of an object in a world as itself a world.) Alluding to the objection that functions are abstract objects while physical objects are not, Cresswell writes:

The view that a function from worlds to worlds is an abstract entity is a question-begging one if being an abstract entity rules out being also a physical object. For on the analysis I am proposing some functions from worlds to worlds are physical objects and so the right conclusion is that not all such functions are abstract entities (p. 97).

One philosophical lesson of our analysis is that these metaphysical uses of set theory are on the wrong track. Neither collections nor types nor worlds nor intensional entities nor physical objects are sets. Many of these things have different identity conditions, and the abstract ones are not objects at all in their own rights, but possible concreta spoken of in an abstract idiom.

The matter of identity conditions is especially important. One suspects, for example, that Cresswell's identification of physical objects with functions harbors
contradiction somewhere, just as there is contradiction in the bold and straightforward identification of sets with collections, as we saw in Chapter 1. But even in that simple case, there was room for enough legalistic maneuvering to evade the contradiction; and Cresswell could, no doubt, escape into a metaphysical briar patch where he would be relatively secure. We shall not attempt to follow him.

There is a legitimate need, of course, to understand physical objects as things which can persist through worlds, or at least through moments of time, and Cresswell is attempting to fill that need by construing them as functions. But there is an alternative which seems to work as well, is in some ways simpler, and is in consonance with our analysis. Worlds, as we suggested earlier, are collections of possible objects. Likewise, moments of time are collections of momentary individuals. An individual extending across time or through worlds is not a function, but merely a collection of these momentary or possible objects. Thus Cresswell is not a function from worlds to worlds or from moments to momentary manifestations, but the collection of all his manifestations in worlds or moments of time. Since collections or physical things are physical, there is no clash of identity conditions here, nor is there any need to identify the abstract with the concrete; persisting physical things and collections of momentary manifestations of physical
things, for example, are both concrete objects.

All of this is not to say, of course, that it is undesirable to represent various objects as sets for formal or mathematical purposes; that is a technique whose legitimacy is beyond doubt. It is only when we become mesmerized by our representations, when we take them for ontological fact, that objections need be raised.
NOTES

CHAPTER 1

1. This, at any rate, is the classical way of understanding unit sets. Quine has a method whereby individuals are identified with their unit sets, though unit sets of nonindividuals remain distinct from their members. See QUINE (1969b), pp. 30-33.

2. See GOODMAN (1951), pp. 44-56, and (1956).


4. See GÖDEL (1944) and STEINER (1973) and (1975a), especially Chapter 4.

5. LEAR (1977) raises some additional epistemological problems for platonistic views generally.


7. The idea is by no means new, though it has only recently received much discussion. Russell himself flirted with it, sometimes regarding his "propositional functions" as open formulas. (For discussions of Russell's notion of a propositional function, see QUINE (1969b), pp. 19 and 241-58, and CHIHARA (1972) and (1973), Chapter I.) Proposals to regard at least some sets as open formulas are also to be found in SELLARS (1963a) and (1963b), PARSONS (1971) and (1974), CHIHARA (1973), and MCDERMOTT (1977). Only the last two present detailed proposals for interpretation of actual set theories. These will be discussed in Chapter 3.

8. Metalinguistic use of set-theoretic concepts is discussed at the end of Chapter 2.

9. Interpretations of transfinite languages like $L^\omega_\Omega$ of Chapter 4 are not functions in the strict sense of sets of ordered pairs, but they may be regarded as classes of ordered pairs. We shall use the word
'function' to designate both sets and classes of this sort.

CHAPTER 2

1. The suggestion of modalizing the quantifiers of set theory has been offered before. PUTNAM (1967) hints at it (pp. 9-12), as does PARSONS (1971), pp. 160-63. CHIHARA (1973) explicitly adopts an approach of this sort (see especially pp. 189-211). We discuss Chihara's view in Chapter 3.

2. It is assumed here that the grammatical rules are consistent. See note 9 of Chapter 6.

CHAPTER 3

1. TARSKI (1936). See also KRIPKE (1975) and the discussion of Chapter 5.

2. Inclusion of P allows us to construct a base language not containing a satisfaction predicate, over whose unary formulas the first metalanguage $L_1$ can be interpreted. Interpretation of $L_1$'s satisfaction predicate cannot begin until we have some formulas whose satisfaction conditions are already defined to serve as its domain.

3. To make the hierarchy impredicative, we could simply drop the qualification "predicative" in the definition of $D_{n+1}$. To make it both predicative and cumulative, we would let $D_{n+1}$ be the set of unary formulas of $L_n$.


5. Ibid., pp. 264-5.


8. Interestingly, however, as is also the case in $\Sigma_\omega$, Cantor's Theorem is unprovable.

9. This can be seen as follows. Let the second-order expansion of a language be the language obtained by adding predicate variables and allowing quantifiers
to bind them. Let the inductive expansion of a language be the result of adding to it all predicates inductively definable from the predicates it contains. It can be shown that every unary formula of the inductive expansion has an extensional equivalent in the second-order expansion and that every formula of the second-order expansion with only an individual variable free has an extensional equivalent in the inductive expansion. Since the standard ways of defining truth for second-order sentences require reference to the class of all subsets (unary predicate extensions) of the domain—a class which cannot be specified constructively for infinite domains—it is likely that truth cannot be defined for the inductive expansion of a language without a similar violation of constructivity.

CHAPTER 4

1. See QUINE (1969b), pp. 19 and 241-58, and CHIHARA (1972) and (1973), Chapter I.


CHAPTER 5

1. For thorough discussion of both systems, see QUINE (1969b), pp. 287-309, or FRAENKEL, et al. (1973), pp. 161-71. NF was introduced in QUINE (1936a) and ML in QUINE (1951), where its power as a foundation for mathematics is demonstrated. ROSSER (1953) gives the development of mathematics from an expanded version of NF.

2. Recall that \( \vdash I(A,f) = T \) for all variable assignments \( f \) such that \( f(x) = a \), and that \( \vdash I(A,f) = F \) for all variable assignments \( f \) such that \( f(x) = a \). Both clauses are needed here because we deal below with nonbivalent systems in which \( I(A,f) \) is sometimes undefined.

3. We could just as well stipulate that interpretations assign three values—T, F and U—which is why the approach can be called "three-valued." But see Kripke's cautionary note on this terminology, KRIPKE (1975), pp. 700-01.
4. For a discussion of STT, see QUINE (1969b), pp. 259-65, or FRAENKEL, et al. (1973), pp. 158-61. MCDERMOTT (1977) also discusses this system and compares it with CTT and ITT.


6. Our definition of NFP below, together with McDermott's work (1977), points to the existence of yet another fragment of NF which may be of some interest—the fragment whose axioms are EXT and all instances of ITT's comprehension principle, omitting superscripts. Just as ITT is intermediate in power between CTT and STT, this new system—call it NFI—is intermediate in power between NFP and NF. NFI is strong enough for the development of a great deal of mathematics, but like ITT it does not permit proof of Cantor's Theorem.

7. One way of approaching this problem which deserves further study is to assume the consistency of NF and then attempt to construct an "inner model" by a procedure similar to that used in Chapter 4 for ZF. Ideally, we should like to prove the existence of an inner model whose "sets" can be construed as stratified formulas of NF and in which $\varepsilon$ is a satisfaction predicate. Such a model would determine an s-saturated fixed point (and would also, of course, be a model of NF). One way to characterize such a model is as follows:

Let $I$ be an interpretation of NF on a domain $D$. Then

(i) $a \in D$ is 0-definable under $I$ if for some unary stratified formula $A(x)$ of the language $L_S$, $\models_I \forall x (x \in y \leftrightarrow A(x))[a]$;

(ii) $a \in D$ is $(n + 1)$-definable under $I$ if for some stratified formula $A(z,x)$ and some $c$ which are $n$-definable under $I$, $\models_I \forall x (x \in y \leftrightarrow A(z,x))[a,c]$;

(iii) $a \in D$ is definable under $I$ if $a$ is $n$-definable under $I$ for some $n \in \omega$.

Let us call a model of NF, each of whose elements is definable, a definable model. A definable model would determine an s-saturated fixed point. For it can be shown by induction on $n$ that all $n$-definable formulas are 0-definable. (The induction makes use of reasoning similar to that used in the proof of LEMMA 3.3.) It follows that each object in the domain has "membership" conditions expressible by some unary stratified formula.
Moreover, for each unary stratified formula $A$, there is some object in the domain whose membership conditions are expressible by $A$, since $\text{COMP}(\text{NF})$ is true. Thus we can in effect identify unary stratified formulas with the objects in the domain whose "membership" conditions (now satisfaction conditions) they express. As a result, $\varepsilon$ becomes a satisfaction predicate, and we have an $s$-saturated fixed point which models NF.

CHAPTER 6

1. Cantor regarded the comprehension principle as self-evident for reasons connected with his views about mathematical existence. He held that a mathematical object exists just in case it can be precisely defined. In the Grundlagen (1883) he writes that "mathematics is, in its development, quite free, and only subject to the self-evident condition that its conceptions are both free from contradiction in themselves and stand in fixed relations, arranged by definitions, to precisely formed and tested conceptions. In particular, in the introduction of new numbers, it is only obligatory to give such definitions of them as will afford such a definiteness...as permits them to be distinguished from one another in given cases. As soon as a number satisfies all these conditions, it can and must be considered as existent and real in mathematics."

(Quoted from the introduction to the Jourdain translation of CANTOR [1895-7], pp. 67-8.)

The idea that precise definition implies existence is reflected again in his definition of the term 'set' (Menge) in the Grundlagen: "By a manifold or aggregate [Menge] I understand generally any multiplicity which can be thought of as one, that is to say, any totality of definite elements which can be bound up into a whole by means of a law" (ibid., p. 54). A law for collecting things into a whole is just a unary formula, and the production of such a formula constitutes the precise definition from which it follows that a set exists. Jourdain interprets Cantor's remarks in the Grundlagen as indicative of a kind of formalist position (ibid., pp. 69-70). Whether or not this is the case, Cantor clearly held that production of a precise open formula was sufficient to establish the existence of a set.

2. They asserted the existence of sets not obviously corresponding to open formulas only when they employed what would now be recognized as the axiom of choice. (The axiom of choice was not explicitly formulated in Cantor's day.) Given an infinite set of sets, it is
not obvious that there is a formula true of exactly one member of each. It would be pointless to speculate about whether the language of naive Cantorian set theory actually contains such "choice formulas," since it cannot consistently be interpreted in the way Cantor thought it could. We know, however, that choice formulas are always available in $L^Z$, since $L^Z$ models the axiom of choice.


4. LEWIS (1973), p. 87. Lewis' argument deals with possible worlds, rather than sets.

5. See RUSSELL (1910), pp. 244-50, and (1919), Chapter XVII.

6. It might be objected that an analogous principle does apply: Do not multiply ambiguities beyond necessity. This objection is quite in order where Tarski hierarchies are regarded as theories about antecedently given data—say, the concepts of truth and satisfaction in a natural language. But in the cases under consideration, they are not theories in this sense at all; they do not serve to account for anything. They are simply artificial creations for dealing with the problem of defining satisfaction for formal languages. As such, they may be elegant or unwieldy, useful or useless, but they cannot violate canons of explanation, since they play no explanatory role.

7. RUSSELL (1908), p. 75.

8. Unless it be evidence that these conventions are consistent—see note 9.

9. He must, of course, have reason to believe that his rules are consistent in order to know that the tokens he is talking about are all possible. In the most favored cases, this can be established by model-theoretic means.

In less favored cases, the best evidence available may be inductive. This is the sort of evidence we have that DEF 4.3 and DEF 4.6, the grammatical conventions for $L^Z$, do not harbor contradiction. We know that if ZF is consistent, tokens of the sort mentioned in DEF 4.3 and DEF 4.6 are possible. And, since a contradiction has never been derived in ZF, despite extensive study, we have good inductive evidence that ZF is
consistent. Therefore, we have good inductive evidence that the tokens of $L^ZF$ are all possible.

In the least favored cases, the grammatical rules governing the object language of a set theory are actually inconsistent. This is the case, as we see it, with naive set theory. We saw at the beginning of Chapter 3 that the attempt to state semantic rules for $L^S$, interpreted as naive set theory, results in both circularity and inconsistency. With inconsistent rules such as this, there is a failure of definition; no range of possible tokens is defined (or the syntax for a range of possible tokens is defined, but the attempted definition fails to provide a semantics). In such a case, of course, there are no possible (interpreted) tokens of the language to know about. (Or perhaps we can know about the possible tokens of a consistent fragment of the interpreted language which we attempted unsuccessfully to define.)

CHAPTER 7


3. Good exposition of the criterion can be found in QUINE (1961), Chapters I and VI, and in QUINE (1969a), Chapter 4.


6. Our thesis applies, however, only to set theory, not to mathematics in general. Whether mathematics as a whole "reduces" to set theory—or whether, if not, analyses similar to ours apply to the unreduced portions—are questions beyond the scope of this study.

7. This is the case, at least, unless much modern linguistic theory is fundamentally misguided. See, for example, DAVIDSON (1967) and Davidson's "Semantics for Natural Languages" or Harman's "Logical Form" in DAVIDSON and HARMAN (1975).

8. See the papers mentioned in note 7.
9. All these examples presuppose that both the language of zoology and the metalanguage in which the rules are stated are English, but nothing would change in substance if they were distinct languages.

10. The extreme conventionalist would deny this. On his view, all statements of science in effect become meaning postulates, so that not only set theory, but the rest of science as well, becomes true by convention. If extreme conventionalism is true, then the distinction drawn here between the role of set-theoretic axioms and the role of the axioms of quantum mechanics collapses. This, however, does not impugn the main thesis of Chapter 7, which is that set theory at least is conventionally true. Indeed, extreme conventionalism entails this thesis.

11. This follows from Gödel-incompleteness. No recursively specifiable set of axioms for ZF is complete; so the standard model cannot be "described" by finite means.

12. Gödel's incompleteness results are sometimes said to show that truth and provability are distinct, so it might be wondered how our analysis can identify them. The answer is that these results imply this conclusion only on the assumption of bivalence. From our point of view, what Gödel proved was only that we are incapable of adopting a set of conventions comprehensive enough to give truth-value to every statement of set theory.

CHAPTER 8

1. This is the case, at least, unless there is some truth to the notion of a language of thought—an intriguing possibility which merits consideration but will not be pursued here.

2. We consider A to be true of \( a_1, \ldots, a_n \) if and only if \( \models A(x_1, \ldots, x_n)[a_1, \ldots, a_n] \), where \( x_1, \ldots, x_n \) are all and only A's free variables in order of first occurrence in A.


4. These remarks are not to be taken as an indication that I share Lewis' realism with respect to possible worlds. I emphatically do not, though I do in a sense agree with him that merely possible worlds are objects of the same kind as the actual one. Merely possible
worlds, as I understand them, are fictions; to spin a
logically consistent yarn and to describe a possible
world are precisely the same thing. They are not
objects to which we are related and about which we
can make empirical discoveries, any more than are
Snow White and the seven dwarves. We know about them
in the same way we know about Snow White—by encounter­
ing or inventing descriptions of them. Yet they are
objects of the same kind as the actual world in just
the sense that Snow White is something of the same kind
as you or me—a person. However, neither Snow White nor
possible worlds exist—though both are possible (vide
the discussion of this terminology in Chapter 2). It is
a corollary of my view that we gain no more philosophi­
cal insight by reconstruing possible worlds as sets
than we do by so reconstruing Snow White.


6. In Cresswell's favor, it should be noted that he
regards his excursions into metaphysics as tentative
and speculative, and that their purpose is primarily
to illustrate in a general way the sort of ontology
required by intensional semantics. Nevertheless, he
insists that his metaphysical proposals are "seriously
intended" (pp. 6–7); so it is not unfair to take them
seriously.
BIBLIOGRAPHY

ALSTON, W. P.

AYER, A. J.

BENACERRAF, P. and H. PUTNAM (eds.)

BERNAYS, P.

BOOLOS, G.

BORGERS, A.

CANTOR, G.

CARNAP, R.

CHIHARA, C. S.
(1968) "Our Ontological Commitment to Universals," Noûs 2, pp. 25-46.


CHIHARA, C. S., Y. LIN and T. SCHAFFTER

CHURCH, A.

COHEN, P.
(1966) Set Theory and the Continuum Hypothesis (W. A. Benjamin, Reading, Mass.).

COHEN, P. and R. HERSCH

CRESSWELL, M. J.
(1973) Logics and Languages (Methuen, London).

DAVIDSON, D.

DAVIDSON, D. and G. HARMAN (eds.)
(1975) The Logic of Grammar (Dickenson, Encino, Calif.).
DAVIDSON, D. and J. HINTIKKA (eds.)


DEVLIN, K.


DRAKE, F.


FRAENKEL, A., Y. BAR-HILLEL and A. LEVY

(1973) Foundations of Set Theory, 2nd ed. (North Holland, Amsterdam).

FRIEDMAN, H.


GÖDEL, K.


GOODMAN, N.


GOODMAN, N. and W. V. O. Quine


KLEENE, S. C.

Kripke, S.

Lear, J.

Lewis, D.

Martens, S.

MCDermott, M.

Montague, R.

Myhill, J. R.
(1952) "The Hypothesis that All Classes are Nameable," Proceedings of the National Academy of Sciences of the United States of America 38, pp. 979-81.

Pap, A.

Parsons, C.

PUTNAM, H.


QUINE, W. V. O.


QUINE, W. V. O. (cont.)


ROSSER, J. B.


RUSSELL, B.


SELLARS, W.


STEINER, M.


TARSKI, A.


VAN FRAASSEN, B. C.


WHITEHEAD, A. N. and B. RUSSELL