Univerzita Karlova v Praze
Filozofická fakulta
Ústav filosofie a religionistiky

Obor: filosofie
Program: filosofie

Karel Procházka

Truth between Syntax and Semantics

Pravdivost mezi syntaxí a sémantikou

Disertační práce

Vedoucí práce: Prof. RNDr. Jaroslav Peregrin, CSc.

2010

Prohlašuji, že jsem disertační práci vykonal samostatně s využitím uvedených pramenů a literatury.

Truth between Syntax and Semantics

Karel Procházka

March 23, 2010

I want to thank my tutor, Prof. RNDr. Jaroslav Peregrin, CSc., for devoting his time and effort to overseeing the present doctoral dissertation, for his critical comments and continuous encouragement. I am also grateful to him for his generous support and his readiness to discuss at length complex philosophical issues. I would also like to thank Doc. PhDr. Vojtěch Kolman, PhD., for reading parts of the text and for his comments and relevant insights.

Abstract

The broad aim of this thesis is to clarify the relationship between syntax and semantics, mainly in connection with languages with exactly specified structure. The main questions we raise are: What is it that makes a semantic concept genuinely semantic? What exactly makes a merely semantic characterization of such a concept inadequate? What is the decisive step we have to make if we want to start speaking about the meaning-side of language? We approach these questions indirectly: via an analysis of a typically semantic concept, namely that of truth. Our principal question then becomes: What conceptual resources are required for a satisfactory definition of truth?

To investigate the concept of truth and different ways in which it can be defined, we have chosen three individual systems: (a cumulative version of) Russell's ramified theory of types, Zermelo's second-order set theory and Carnap's logical syntax. Each of the systems is studied in considerable detail. The presented thesis is, in effect, a collection of three case-studies into the ways in which the concept of truth is explicitly definable and into the requisite conceptual background, each study forming a more or less closed unity. It should be noted that we are not interested in a historically faithful representation of these systems; our goal is to get the best of them while making use of suitable contemporary insights.

The general conclusion reached on the basis of the results obtained in the individual studies of the concept of truth is that the key step marking the transition from syntax to semantics consists in a specific combination of restricting and expanding the syntactic resources available. At the very end, some philosophical consequences following from this idea are outlined.

Abstrakt

Širším cílem této práce je vyjasnit vztah mezi syntaxí a sémantikou, zejména pokud jde o jazyky s přesně specifikovanou strukturou. Hlavní otázky, kterými se zabýváme, jsou: Co činí sémantický pojem sémantickým? Co způsobuje, že je pouhá sémantická analýza takového pojmu nedostatečná? Co je tím rozhodujícím krokem, který musíme učinit, abychom pronikli k významové stránce jazyka? Těmito otázkami se nezabýváme přímo, ale prostřednictvím analýzy typického sémantického pojmu, a sice pravdivosti. Naší hlavní otázkou tedy je: Jaké pojmové prostředky jsou nezbytné pro uspokojivou definici pravdivosti?

Ke zkoumání pojmu pravdivosti a jednotlivých způsobů, jak jej lze definovat, jsme si vybrali tři konkrétní systémy: kumulativní verzi Russellovy rozvětvené teorie typů, Zermelovu druhořádovou teorii množin a Carnapovu logickou syntax. Každý systém je podroben důkladnému studiu. Předkládaná práce je tedy souborem tří více méně samostatných studií, jež popisují možnosti explicitní definice pravdivosti a nezbytného pojmového zázemí. Poznamenejme, že naším cílem není historicky věrná prezentace uvedených systémů, nýbrž snaha o další rozvinutí toho cenného, co nabízejí, ve světle současných poznatků.

Obecným závěrem, k němuž dospějeme na základě výsledků získaných v jednotlivých studiích pojmu pravdivosti, je teze, že klíčový krok, který je třeba učinit pro přechod od syntaxe k sémantice, spočívá ve specifické kombinaci omezení a rozšíření syntaktických prostředků, které jsou k dispozici. Na samotný závěr načrtneme některé filosofické důsledky tohoto zjištění.

Contents

Chapter 1

Introduction

This thesis deals with four central concepts: syntax, semantic, truth and meaning, and with their mutual relationship. Although none of these concepts is regarded as more eminent than the others, they are not treated in an equal way. This follows from the consideration of how they are interrelated. Syntax and semantics represent—alongside the third member of the "trinity", pragmatics—two distinct layers or aspects of language. Semantics as a discipline is usually characterized as a study of all the various aspects of meaning of linguistic expressions. This, however, does not entail that to study semantics, we have to be directly occupied with the concept of meaning. The way the concept of meaning is approached in this thesis is inspired by the idea that in order to contribute to the study of the meaning-side of language, it is not necessary to invoke the concept of meaning at all.[1] Semantics is not an enquiry into the unique concept of meaning. Rather, it strives to account for how it comes about that expressions mean what they do. Thus we can contribute to the study of meaning by elucidating the functioning of some particular concept other than meaning itself. If we succeed in elucidating the meaning of a particular word, if we gain a better understanding of how a complex concept functions in our language, we have solved one problem of semantics, and we have clarified a little region in the abstract realm of meaning.

The concept we have taken up as a focal point of our investigations is that of truth. There is a sense in which truth cannot be appropriated by any single science or domain of knowledge since they all are after the truth; however, the questions such as: What makes a sentence or a statement true? What is the general criterion of truth? How can truth be defined? are usually treated within semantics. Yet, how about syntax? What does syntax has to do with meaning or truth, these being chief topics of semantics? The key question we are going to ask can be formulated in the following way.

[1] To be more precise, we are inspired by Dummett's depiction of the theory of meaning in Dummett [1981], p. 675.

By definition, the syntactic components of language taken in isolation are meaningless and there is no way how we could attribute to them semantic properties such as truth. To get hold of the meaning or truth of expressions, we need to take into account the whole semantic aspect of language. But what is that? How do we add the semantic dimension to language? How do we assign meanings to expressions or what do we do to gain the ability to treat them as true or false? What is it that marks the transition from syntax to semantics? When do we stop doing syntax and begin doing semantics? What is the fundamental move that could be identified with the passage to semantics?

Our main task is to answer this complex of questions. First of all, though, we need to make the whole issue more precise.

1.1 Syntax and Semantics

It needs to be said right at the beginning that we will deal mainly with formal languages. (This decision is further discussed in section 1.4.) Unlike in so-called 'natural languages' such as English or Czech, in formal languages it is precisely specified what counts as a well-formed expression of such a language, be it simple or complex (e.g., a sentence). This is usually achieved by specifying the basic symbols and the rules of formation of the given language. This part of language, i.e., the vocabulary and the formation rules, constitutes syntax in the narrow sense; it is, in effect, combinatorics of some chosen basic elements. Sometimes also the formal specification of what counts as a theorem is included in syntax. This is typically done by laying down axioms and rules of inference; no reference to an intended meaning is yet made. This is syntax in the wider sense. A full exploitation of this deductive component of language leads to proof theory. The deductive component is often constituted not only by logic but also by axioms or rules of special theories. A clear-cut separation of logic from other, extended systems is not always unproblematic and can sometimes be viewed as more or less arbitrary. We will generally stick to the convention of regarding standard predicate calculus, both first- and higher-order, as formal logic while collections of axioms including other primitive expressions as formal theories.

So far we have considered only syntax. Semantics begins when we start to enquire into the interpretation of the language, so far considered purely formally. We investigate the meanings of expressions, truth or falsity of sentences, definability of concepts by formulas etc. Typically, the field of study dealing with interpretations of formal systems is model theory.

As an illustration of a formal theory, take Peano Arithmetic, PA. It is formulated in a standard language of first-order logic, to which it adds the constant '0', the one-place functor 's' and the two-place functors '+',

'×'. Its formation rules are standard. All this belongs to syntax in the narrow sense. Its deductive system consists in standard axioms and rules of inference of first-order logic plus the six axioms governing the use of the aforementioned extra-logical constants and the axiom schema of induction; it gives us properties such as that of being a PA proof or being a theorem of PA. This constitutes syntax in the wide sense. Semantics for PA is typically provided by specifying a structure (model) which interprets the parameters of PA, i.e., the quantifiers and the extra-logical constants listed above, and which makes all theorems of PA true. The structure $\langle \mathbb{N}, +, \times, 0, s \rangle$—where '$\mathbb{N}$' specifies that the quantifiers of the language of PA range over the domain of natural numbers, and the remaining symbols assign individuals from $\mathbb{N}$ and functions on $\mathbb{N}$ to the constants—is taken to be the standard model of PA. Consideration of the intended model and other possible structures satisfying PA leads to the development of a variety of semantic concepts for the given formal theory.

It should be stressed that without an interpretation, a formal language is only a collection of basic symbols from which we or a suitably designed machine can generate, following explicit rules, well-formed formulas or other kinds of expressions. Similarly, a formal theory is a collection of sentences that are generated from some initial sentences we accept from the start, again mechanically applying explicit inferential rules. In this sense, a formal theory is just a collection of specific objects, say, strings of symbols. As it is not interpreted, it cannot be seen as saying or asserting anything at all. This changes when we consider a formal language together with its interpretation, i.e., when we take into account how the strings of symbols are to be "read". It is clear that unless it is provided with semantics, a formal language is not a genuine language at all. So to establish a formal language in the full sense is to provide both its syntax and its semantics.[2]

Before it took on the respectable, technical form of model theory, semantics was regarded by "scientifically-minded philosophers" with suspicion. The reason is not hard to find. The assignment of meaning to a given formal language is seen as establishing a relationship between the expressions of a language and something extra-linguistic. That is, in dealing with the semantics of a language, we leave the realm of the linguistic and enter the world of objects considered outside the medium in which they are given to us. This is, indeed, problematic on several accounts. Firstly, it does not seem, at least at a first glance, unacceptable to say, e.g, that the phrase 'the tree outside the window' designates the tree growing outside the window. However, when we leave the domain of discrete tangible objects, the things get complicated. With Frege, we may raise the question: What is the meaning of a number

[2]Sometimes the term 'formal language' is used only to refer to the uninterpreted syntactic part; at times the distinction between the *formal* language (excluding interpretation) and the *formalized* language (including both syntax and semantics) is made. We do not observe this distinction.

word? In general, what are the assertions involving designations of abstract objects about? What about the talk of properties and relations? To provide an acceptable interpretation of a more complex language is nothing but to answer some perennial philosophical questions that keep haunting us from the antiquity. The complexity of such a task only grew with the discovery of paradoxes at the beginning of the 20th century. Secondly, it must be asked: How do we grasp the extra-linguistic entities that are assigned to linguistic expressions as their meanings? In the same way as it is hard to find out what the object that is seen *looks like* when it is not seen, it is also hard to determine what the meaning expressed by a sentence is when it is not expressed by any sentence. The task of interpreting a language seems to involve this hardly acceptable requirement that we should be able to grasp the meanings we assign to expressions in some non-mediated way.

Semantics responds to this difficulty by strictly observing the distinction between the object-language and the metalanguage.[3] The interpretation of expressions of the object-language is not carried out in an extraterrestrial observatory but simply in a metalanguage. The meanings to be assigned are expressed or designated by the expressions of the metalanguage. This is to say that the object-language is interpreted in the metalanguage. In practice, this will typically include a translation of the object-language into the metalanguage, as a result of which we will be able to associate the expressions of the object-language with the entities designated by the metalanguage.

This is all fair enough; but the necessary requirement is that the metalanguage itself is already interpreted, i.e., it is a fully-fledged language with proper semantics. Of course, one can formalize also the metalanguage but this just pushes the requirement that we be in possession of a fully meaningful language as a metalanguage one step further. In the end, if we are to avoid an infinite regress, we end up with a language for which we do not have any metalanguage. How does one study the semantics of such a language? With no metalanguage to climb upon and with the assumption that the option of acquiring a direct acquaintance with the meanings is hardly acceptable, this questions does not seem to have an easy answer.

1.2 Truth

To make our investigations more specific and as focused as possible, we have picked out a single central semantic concept, namely that of truth. The significance of the property of truth for formal theories became apparent in connection with Gödel's incompleteness theorems. Before Gödel's groundbreaking result it could have been hoped that a suitable choice of axioms together with proper rules of inference would make it possible to prove

[3] Cf. Tarski [1936*b*], p. 402: 'People have not been aware that the language *about which* we speak need by no means coincide with the language *in which* we speak.'

every true sentence of the given language. Had this been the case, truth would turn out to coincide with theoremhood: truth would be whatever a well-constructed formal theory can prove. Then, of course, truth would be redundant and could be eliminated without a significant loss. Yet, this was not meant to be. The incompleteness theorems changed profoundly our understanding of formal systems. Among other things they established that provability and truth do not generally coincide. It follows that truth can be and should be investigated as a unique, independent concept in its own right, and that the study of the truth-aspects of formal languages is a genuinely worthwhile task that duly supplements the study of their syntactic aspects.

The conception of truth usually associated with the semantic study of language is that of Tarski's, published and promoted in the 1930s. This is not to say that truth was not studied before Tarski but it was in Tarski's monograph (Tarski [1933]) where a precise axiomatic treatment and an explicit definition for truth was proposed for the first time.[4] Without going into details (as these are provided in subsequent chapters, especially throughout chapter 4), it suffices to say that Tarski proceeds in two steps. First, he introduces a criterion specifying what is to count as a truth predicate for a given object-language. This is nothing else than the well-known convention T. The second step consists in giving the definition of truth based on another semantic notion, namely on the relation of satisfaction of formulas by sequences of objects. The specifications of satisfaction and of truth can take on two different forms, depending on the expressive or deductive resources available in the metalanguage. If the metalanguage is what Tarski calls "essentially richer" than the object-language, the relations of satisfaction and truth can be explicitly defined. This means that they can be eliminated, i.e., replaced in the metalanguage by their definienda, and that the metalanguage does not need to contain expressions such as 'true' or 'satisfies' among its primitive vocabulary. On the other hand, if the metalanguage is not essentially richer than the object-language, the explicit definition is not possible. Satisfaction and truth can then be defined at most implicitly, i.e., the predicates 'true' and 'satisfies' must belong to the primitive vocabulary of the metalanguage, and the deductive system of the metalanguage (the metatheory) has to include special axioms governing the use of these undefined constants. The specification of truth in this latter case will thus have a form of an axiomatic theory of truth, which is in no way eliminable.

So Tarski's semantic conception of truth offers two possibilities. On the first, truth for the object-language can be fully subsumed under the essentially richer conceptual machinery of the metalanguage. On the second, it can be introduced via suitable axioms as an undefined primitive concept. Both alternatives have their philosophical drawbacks. In the former case,

[4]For a slightly more complete account of the story see the beginning of chapter 5.

we already need to assume what we are looking for or even more, only at a higher level. In the latter case, the axiomatic theory of truth is, in effect, a formal theory which can be understood in terms of syntax in the wider sense; as such it requires an interpretation. That is, we need to assign meaning to the primitive constants 'true' and 'satisfies'. Now, in view of these drawbacks, we do not mean to suggest that the semantic conception of truth is without merits. On the contrary, it led to the rehabilitation of the concept of truth in logic and mathematics, it opened up the possibility for this concept to be employed in a consistent fashion, and it brought out the compositional character of the relation of satisfaction, etc. However, it can hardly be seen as an ultimate solution to the philosophical problem of grasping the meaning-side of language.

In the light of this somewhat disappointing situation, is there a more promising way of approaching truth other than either via accepting the existence of a peculiar cognitive capacity that would show us the absolute truth in itself or via assuming the same concept that is to be clarified in the metalanguage? Another question closely related to the one just asked is this: Is there a way in which the truth-predicate could be introduced and consistently used without climbing up onto the appropriate metalanguage? That is, is the transition to a different language really a necessary condition for an acceptable definition of truth?

1.3 The Problem and the Way Forward

The chief problem we aim to deal with can be formulated in the following way. Assume that truth can be somehow defined for a particular language. Given that it is a semantic concept par excellence, i.e., a concept that has to do with the meaning-side of language and cannot be properly exploited using syntactic means alone, its definition will necessarily involve certain semantic elements. What are these semantic elements that such a definition must make use of? This is a crucial question. Answering it will provide, among other things, an answer to the question with which we have started, namely that of what marks the passage from syntax to semantics, at least with respect to this specific concept. A particular task is to find out what role in such a definition of truth is played by the distinction between the object-language and the metalanguage.

It is hoped that if we manage to obtain a satisfactory answer to these questions, it will become clearer whether the concerns raised in connection with the very business of semantics can be dispelled, and, if this is so, how it can be done. There is a chance that a thorough study of the elements that make the concept of truth definable—i.e., of the elements to which truth can be reduced or of the components of which it is built—will open up a different way of looking at the modus operandi of semantics and suggest a somewhat

different, philosophically less discouraging approach to the whole meaning-side of language. This is, indeed, a very ambitious project, requiring a great deal of hard labour without any guarantee that the results actually obtained will be worthwhile.

The strategy adopted for dealing with the task we have set for ourselves consists in carrying out three more or less extensive case-studies into the different ways in which the concept of truth is definable in particular systems. The method chosen is partly quasi-historical, and partly systematic. I have used the term 'quasi-historical' to emphasize that we are going to deal with systems put forward by actual philosophers and logicians within determinate historical contexts but we do not at all attempt to present them with historical accuracy. The historical systems are taken mainly as material for a further study, which is conducted with great liberty. That is, we attempt to extract from our material as much as possible, and we sometimes develop the systems into shapes that they historically never assumed. The systematic element suggests that we strive to make use of the most contemporary theoretical resources and analyze the concepts using modern techniques; sometimes, as in chapter 4, we focus on the systematic development entirely, leaving historical considerations almost entirely aside.

To be specific, we investigate how truth is defined within the following systems: (a cumulative version of) Russell's ramified theory of types, Zermelo's second-order set theory and Carnap's logical syntax. Each case-study comprises a single chapter except for the second one which spans two chapters, the latter being concerned not so much with Zermelo's thoughts but with prospects for the definition of truth in set theory in general. This is the most technical part of the whole thesis, and some ideas and techniques only assumed or informally touched upon in other sections are fully developed there. The individual studies are, on the whole, supposed to be independent of one another and should form more or less integral wholes. They may be read in a different order without seriously impairing the reader's ability to follow the arguments presented and concepts discussed.

In all the three studies, the focus is on the development of the concepts needed for the definition of truth. The problem formulated at the outset of the present section is not explicitly addressed until the conclusion, in which an attempt is made to discern a unifying thread in the different theories and approaches examined in the body of the thesis, and conclusions are drawn.

1.4 Formal Languages

At the beginning of section 1.1 we made the decision to deal with formal languages, in the sense specified, and not with natural languages. Formal languages are constructed languages; their rules have been explicitly stated, which makes them unproblematic in some areas where natural languages,

owing to their complexity and instability, are rather difficult to come to terms with. Does not this decision threaten to undermine the validity of our whole project? What philosophical impact can have a definition of a technical concept whose use is entirely restricted to a specific constructed language? How does it contribute to our grasping of the concept of truth for the language we commonly speak if we succeed in obtaining a precise explicatum for truth, say, in second-order set theory? The question of the relationship between constructed and natural languages and the philosophical relevance of the former for understanding the latter is a rather difficult one. We will not attempt to solve it. We will merely make a few remarks that will hopefully bring some light into the issue and will, in the end, justify our decision.

Formal languages and formal theories can be understood in diametrically opposite ways. On the one hand, they can simply be objects of study (thence 'object-languages'). This approach is decisively external: we investigate the syntax and the semantics of a formal language from the outside, from the vantage point of a different language, which is no longer studied as an object but actively used. Taken as objects of study, formal languages or theories can be viewed as models; not in the sense of models of model theory but as simplified representations of selected aspects of reality that are put together in such a way that their nature and mutual relations become more transparent and detectable. On the other hand, formal languages and theories can be seen as fragments of natural languages. Let us look at this option more closely.

When Frege introduced his formal language called 'Begriffsschrift' in Frege [1879], he considered the question of how it is related to natural language.[5] Frege's formal language arose out of the need to overcome the inadequacy of natural language for the purpose of reliably treating the chains of inference, i.e., proofs. Frege compares the usefulness of his formal language to that of a microscope for the human eye. For general purposes of seeing the world around us, the eye alone is more useful; however, for a very specific sort of purposes, its capacities need to be extended. Similarly for the proposed formal language: its utility consists in analyzing proofs and logical connections between sentences, which are not always manifest in natural languages.

However, since logical relations can be seen as underlying any meaningful use of language, the actual ambitions behind Frege's formal language are far greater:

> It is possible to view the signs of arithmetic, geometry, and chemistry as realizations, for specific fields, of Leibniz's idea [of a universal characteristic]. The ideography proposed here adds a new

[5] These considerations are presented in the preface to *Begriffsschrift*. Viz. Frege [1879], pp. iii–viii.

> one to these fields, indeed the central one, which borders on all the others. If we take our departure from there, we can with the greatest expectation of success proceed to fill the gaps in the existing formula languages, connect their hitherto separated fields into a single domain, and extend this domain to include fields that up to now have lacked such a language. (Frege [1879], p. vi; Bauer-Mengelberg's translation.)

Then Frege goes on to suggest that his formal language may be extended so that it applies to analysis, to geometry and to physics, and there is no obstacle to including other scientific disciplines. This is to say that Frege's formal language is proposed as a formal language that can incorporate, with suitable extensions, the whole language of science. And not only that; the ambition connected with it goes even further. Being a 'formula language of pure thought', it can include, contrary to Kant's strongly held belief,[6] also the field of philosophy:

> If it is one of the tasks of philosophy to break the domination of the word over the human spirit by laying bare the misconceptions that through the use of language often almost unavoidably arise concerning the relations between concepts and by freeing thought from that with which only the means of expression of ordinary language, constituted as they are, saddle it, then my ideography, further developed for these purposes, can become a useful tool for the philosopher. (Frege [1879], p. vi–vii; Bauer-Mengelberg's translation.)

It follows from these considerations that Frege's formal language was not thought of as an object of study. It was put forward as a language into which sentences of natural language as well as sentences belonging to other possible formal languages of different special sciences are translatable, and its whole raison d'être was to make explicit the logical form of these sentences. It was designed to be a language that is to be actively used, and not just studied as a representative of a rare species. It makes more sense to look upon it as a precisely crafted fragment of natural language rather than as a remote system of formal logic, separated by a deep divide from the rest of our linguistic abilities and practices.

There is no doubt that Russell's ramified theory of types was conceived of as an attempt to establish a universal logical language in which the somewhat hidden logical form of sentences of natural language could be exposed. As far as Zermelo's system of set theory goes, it makes sense to see it as

[6] Kant consistently refused the idea that philosophy could successfully employ mathematical methods such as the use of symbolism. The reason is that while mathematics relies on intuition, philosophy must deal with pure concepts. See Kant [1998*b*], A712/B740–A738/B766, in particular A734/B762–A735/B763.

a precisely specified fragment of the language of science whose chief goal is to fully exploit the mathematics of the membership relation. With Carnap's syntax, the situation is a bit more complex as Carnap programatically refuses to accept that there is a single language of science, claiming that everyone has freedom to choose the language he or she finds best suited for the desired purposes. This perspective of language plurality involves consideration of the individual candidate languages as objects. However, Carnap is sensitive to the distinction between working within a language and studying a language from the outside. The languages proposed as candidates for playing the role of language of science need to be studied so that we are aware of their powers and limitations as well as their advantages or disadvantages in comparison with other candidates. However, the ultimate goal of any such candidate language is to become a language of science, i.e., an explicitly specified language that is to be used to advance progress of all scientific disciplines. In this sense, Carnap's candidate languages should not be thought of as aspiring at being accepted as suitably sharpened fragments of natural language; they are put forward as its adequate replacements. They are to become natural languages sui generis.

To conclude, there is a perspective that permits viewing formal languages as languages that can be used. This makes them genuinely philosophically significant.

Chapter 2

Russell: Hierarchy of Functions

The main aim of this chapter is to show how truth can be defined in Russell's ramified theory of types. Yet it needs to be noted right at the beginning that Russell's brisk development combined with not so clear intentions behind some of his assertions make it at some points rather difficult to estimate what his true philosophical positions were. Fortunately, as what we are after is not a historical investigation but rather a further development of certain possibilities contained in the ramified theory of types, we are free to make use of a modern reconstruction of the theory of types and a particular construal of the underlying hierarchy of propositional functions without much regard to the question whether it can be genuinely ascribed to Russell.

It is well known that the theory of types was not something Russell was ecstatic about; it arose out of the necessity to save the foundations of logic and mathematics from contradictions such as those that afflicted the system developed in Frege's *Grundgesetze der Arithmetik* (Frege [1893, 1903]). So, before turning our attention to Russell's diagnosis and the proposed cure, it is requisite to understand the problem, and to say a few words about the collapse of Frege's attempt to coach arithmetic in logical terms. After this preparatory part, we will describe Russell's response, the vicious-circle principle and the ramified theory of types. Then we will show how semantic concepts such as truth can be defined within our broadly Russellian system. Eventually, we will tackle the issue of expressibility within the language of the ramified theory of types.

2.1 Frege: Trouble with Basic Law V

Frege's project aiming to show that arithmetic is in essence nothing but logic is well known and has been widely discussed, especially in connection with different sorts of 'neologicist' attempts to partially revive the basic

idea without giving rise to a contradiction or to a commitment to blatantly unlogical assumptions.[1] In the light of this, we will not describe Frege's conception of logic and arithmetic in any detail; we will just say the following. Arithmetic deals with numbers. Numbers are understood as unique objects given to us in language as the meanings of number-words. The context principle requires that we investigate the meaning of a number-word in the context of sentences containing it. A prominent role for determining the meaning of a word is played by sentences expressing the identity of objects. Frege's task then becomes to determine the meanings of the number-words flanking the equality sign '=' and secure their existence in such a way that the sentences of arithmetic can be assigned a definite truth value. The True and the False are assumed as primitive objects, and the other members populating the domain of objects required for the sentences of arithmetic to be true are provided by specifying the meanings of complex terms in identity statements. Where do these objects come from? What are they, and in what sense can they be said to be logical? The crucial idea of Frege's is that these objects are constituted as extensions of concepts, or, more generally, courses of values, or value-ranges, of functions.

Frege considers the logical universe to comprise two basic kinds of entities, functions and objects, which are separated by an unbridgeable divide.[2] However, despite being essentially different from objects, functions are thought to be associable with certain objects, namely with the courses of their values. A concept, a special case of function whose value is a truth value, is associated with its extension, i.e., the collection of objects falling under it. To designate functions, Frege replaces the constants in the argument position by Greek consonants representing the empty argument place, e.g., '$f(\xi)$'. A course-of-values of a function is represented by using a Greek vowel and prefixing a binding operator, which is the same Greek vowel with the smooth breathing, i.e., '$\grave{\epsilon}f(\epsilon)$' designates the course-of-values of a function $f(\epsilon)$, and '$\grave{\epsilon}F(\epsilon)$' the extension of a concept $F(\xi)$.

The principle that introduces courses-of-values into Frege's system is nothing else than the infamous Basic Law V:

$$\grave{\epsilon}f(\epsilon) = \grave{\alpha}g(\alpha) \leftrightarrow \forall x(f(x) = g(x)), \qquad \text{(A-BLV)}$$

which says that a function $f(\xi)$ has the same course-of-values as a function $g(\xi)$ if and only if both of the functions always assign the same value to the

[1] For a nice survey of basic neologicist strategies, see Linsky and Zalta [2006].

[2] The key characteristic separating functions from objects is, according to Frege, the fact that the function considered in itself is not a complete whole, or is 'unsaturated' (cf. Frege [1891], p. 6). The object is specified negatively: it is anything that is not a function (op. cit., p. 18). It is important to realize that the distinction strictly complies with the context principle in the sense that whether a string of words designates a function (or a concept) or an object depends on its function within a sentence. Thus, to use an often quoted example of Frege's, 'the concept *horse* is not a concept' but an object (Frege [1892], p. 196).

same argument.[3] Obviously, Basic Law V codifies the criterion of identity for courses-of-values. However, it is actually much stronger than that; it has important corollaries that reveal that it is a fully-fledged abstraction principle. Using (A-BLV), it is easy to prove that every function has a course-of-values:

$$\forall f \exists x (x = \grave{\epsilon} f(\epsilon)).$$

(To prove this, we instantiate f to g in (A-BLV), then, as $\forall x(g(x) = g(x))$ is valid, we derive that $\grave{\epsilon} g(\epsilon) = \grave{\epsilon} g(\epsilon)$, from which it follows by existential generalization that $\exists x(x = \grave{\epsilon} g(\epsilon))$, which can be universally generalized on g to obtain the corollary.) Furthermore, given that the membership relation is explicitly definable in Frege's system, (A-BLV) entails the principle of extensionality as well as the classical "naive" comprehension principle: $\forall F \forall x(x \in \grave{\epsilon} F(\epsilon) \leftrightarrow F(x))$.[4] So Basic Law V is a very powerful principle establishing not only the criterion of identity but also the existence of a special kind of abstract objects together with some of their basic properties.

Of course, it is well known that Frege's system is inconsistent, and that Basic Law V represents the "serpent of inconsistency" in Frege's paradise.[5] In order to see where the trouble lies, it is important (and almost sufficient) to realize that—as the course-of-values $\grave{\epsilon} f(\epsilon)$ is an *object* which corresponds to a first-level *function* $f(\xi)$—Basic Law V, in effect, posits a second-level function that maps first-level functions to objects.[6] This fact can be expressed by the following statement:

$$\exists m \forall x \forall g \forall f \big(m_\beta(f(\beta)) = m_\beta(g(\beta)) \leftrightarrow f(x) = g(x) \big), \qquad \text{(S-AlB)}$$

in which the sign '$m_\beta(f(\beta))$' is Frege's expression for a second-level function. A half of this biconditional does not lead to problems, namely the implication from the right to the left, which says that there is a second-level function m that yields identical values whenever its arguments (i.e., first-level functions) return the same values for the same arguments. In other words, there are no two coextensive first-level functions for which the second-level function m would return two distinct values. The implication that is responsible for

[3] Cf. Frege [1893], §3. If we wanted to use a more contemporary notation, we would need to distinguish between Basic Law V as applicable to functions (using Church's lambda-notation):

$$\lambda x f(x) = \lambda x g(x) \leftrightarrow \forall x(f(x) = g(x))$$

and Basic Law V as applicable to classes:

$$\{x \mid F(x)\} = \{y \mid G(y)\} \leftrightarrow \forall x(F(x) \leftrightarrow G(x)).$$

Cf. Kolman [2002], p. 201. In Frege's system, though, as classes are extensions of concepts, and concepts are nothing but special cases of functions, the class version would be seen merely as a more specific variant of the version for functions.

[4] For proofs, see Zalta [2009], section 2.4.

[5] Cf. Dummett [1991], p. 209.

[6] Cf. Ricketts [1997], p. 199.

the contradiction is the one from the left to the right, which says that there is a second-level function m which is such that whenever it yields identical values, its arguments (i.e., first-level functions) are coextensive. To put it differently, the faulty implication states that our second-level function has to be injective, i.e., that it is not the case that one value yielded by the second-level function m gets associated with two non-coextensive first-level functions.

The trouble lies in the fact that in Frege's system it can be proved that there is no second-level function of which the statement (S-AlB) would be true. Frege himself proves this result in the appendix to the second volume of *Grundgesetze*[7] by deriving the following theorem (concerning concepts, though, not functions in general):

$$\forall m \exists x \exists G \exists F \big(m_\beta(F(\beta)) = m_\beta(G(\beta)) \wedge \neg [F(x) \leftrightarrow G(x)] \big). \qquad \text{(S-}\chi\text{)}$$

This theorem says that for every second-level function (whose only argument is a first-level function of one argument) there are first-level concepts to which it assigns the same values although these concepts are not coextensive. To put it simply, there is no injective function from concepts to objects. As Frege executed the derivation of the theorem (S-χ) without the use of Basic Law V, he showed that the rest of the logical system of *Grundgesetze* is incompatible with it.

Basic Law V thus turns out not to be a law, i.e., a truth that is universally valid. It follows from (S-χ) that there are certain "rogue" concepts that block its universality. What do they look like? Russell's paradox, which revealed inconsistency of Frege's system, involved the property of being a class that does not belong to itself, or, as Frege puts it, the property of being the extension of a concept that does not fall under this very concept. In Frege's system, this property can be expressed as follows:

$$R(\xi) \leftrightarrow_{Def.} \exists G(\grave{\epsilon} G(\epsilon) = \xi \wedge \neg G(\xi)). \qquad \text{(D-R)}$$

It is easy to see that $\grave{\alpha} R(\alpha)$ has the property R, i.e., that $R(\grave{\alpha} R(\alpha))$. Yet if we expand R back into its full form, we obtain the following: $\exists G(\grave{\epsilon} G(\epsilon) = \grave{\alpha} R(\alpha) \wedge \neg G(\grave{\alpha} R(\alpha)))$. This result, combined with $R(\grave{\alpha} R(\alpha))$, immediately yields a counterexample to Basic Law V:

$$\exists G \big(\grave{\epsilon} G(\epsilon) = \grave{\alpha} R(\alpha) \wedge [\neg G(\grave{\alpha} R(\alpha)) \wedge R(\grave{\alpha} R(\alpha))] \big). \qquad \text{(S-Coe)}$$

(There is a variant of this argument, also considered by Frege, based on the property $\forall G(\grave{\epsilon} G(\epsilon) = \xi \rightarrow \neg G(\xi)))$.[8]

[7] Cf. Frege [1903], p. 260. The appendix was hastily added after Frege was informed of Russell's paradox. See also Kolman [2002], p. 232.

[8] However, what we have constructed is not, properly speaking, an explicit counterexample to Basic Law V but merely a statement asserting its existence. Frege acknowledged

To sum up, the cause of the contradiction lies in the "brute fact" that the power set of a set cannot be injected into that set,[9] which result is otherwise known as 'Cantor's theorem'. There are always strictly fewer objects than (first-level) concepts, so there cannot be any injective function from concepts to objects, neither can there be any surjective function from objects to concepts.

2.2 From Concepts to Objects

The introduction of courses-of-values as abstract objects associated with functions was an essential element of Frege's effort to provide arithmetic with a logical foundation. He was rather clear about this:

> I should gladly have relinquished [Basic Law V] if I had known of any substitute for it. And even now I do not see how arithmetic can be scientifically founded, how numbers can be conceived as logical objects and brought under study, unless we are allowed—at least conditionally—the transition from a concept to its extension. (Frege [1903], p. 253; Furth's translation in Frege [1964], p. 127.)

Moreover, Frege realized that what was under a threat was not only the system he had laid down in *Grundgesetze* but any system making use of extensions of concepts, classes or sets and, in general, the whole project of establishing a logical foundation for arithmetic.[10] So it seems that the logicist project can survive only if one manages to save logical objects. However,

---

that he was unable to explicitly define such a concept itself:

> We should like to have an example of this: how is such a concept to be found? It cannot be done without a more precise specification of our function $\dot{\epsilon}\Phi(\epsilon)$, of the extension of a concept; for our previous criterion for the coinciding of extensions at this point forsakes us. (Frege [1903], p. 262; Furth's translation in Frege [1964], p. 138.)

Thus unlike the well known Cantorian argument against the existence of a second-level function from objects onto first-level functions—which provides the explicit definition of a "rogue" first-level function that does not get associated with any object—the Fregean argument against the existence of an injective second-level function from first-level functions to objects described above offers just a general existence claim. (The same goes for the aforementioned variant of the argument.) It is not the case, though, that an explicit definition of a counterexample to Basic Law V cannot be constructed. Two versions of such an explicit definition can be found in Boolos [1997].

[9] Cf. Boolos [1993], p. 234.

[10] In a footnote (Frege [1903], p. 253), Frege picks out Dedekind and his 'systems' as an example of such an endangered species. Indeed, as Kamareddine *et al.* [2002], p. 199, point out, Frege's mature system of *Grundgesetze* was not the only inconsistent system out there since Russell's paradox 'could be formulated in all the systems that were presented at the end of the 19th century (except for Frege's *Begriffsschrift*)', in particular in the influential and popular system of Peano's. The case of Cantor is discussed below.

what are the chances of success? If Basic Law V has failed, what can replace it? Is it really possible to establish the existence of abstract objects using logic alone?[11]

Frege responded to Russell's paradox by providing a quick patch to Basic Law V which excluded from its scope precisely those "rogue" concepts that were identified as responsible for the contradiction. Nevertheless, not only was the amended principle unable to serve as a proper abstraction principle but it fared no better that the original Basic Law V. The amended system was proved to be contradictory by Leśniewski[12] and Quine [1955].[13] As Potter points out, Russell's paradox 'is merely the simplest of a great variety of paradoxes which can be derived in Frege's system; blocking one is no guarantee of blocking them all' (Potter [2000], p. 113). To be just, it should be admitted that Frege well understood that his system was in ruins, and without any doubt did not expect anything from his quick patch.[14] Later on, in a letter written in 1925, Frege exclaims: 'We must set up a warning sign visible from afar: let no one imagine that he can transform a concept into an object' (Frege [1980], p. 55). The transition from concepts to objects is declared impassable.

Still, before pronouncing the final verdict, Frege outlined two different general strategies for blocking Russell's contradiction while retaining extensions of concepts as objects. The first strategy is based on the denial of the universality of the abstraction principle. That is, it is conceded that some functions are not associated with any courses-of-values. Frege rejects this path since it forces us to accept that there are properly formed symbols in our language—viz. the names of the courses-of-values—that do not denote anything. The sham names of courses-of-values cannot then be seen as dissociable logical components of meaningful larger compounds,[15] therefore, it will be illicit to replace them by variables. But then, as numbers are

[11]Note that we are not exhausting all options. One might give up the idea of obtaining a domain of well-established abstract objects that could count as logical, and one might argue that a form of the logicist program is still achievable if we follow the strategy developed and rejected by Frege in his *Grundlagen der Arithmetik* (Frege [1884]) that is based on a different abstraction principle, called 'Hume's Principle'. This is the neologicist position promoted, above all, by Crispin Wright. For arguments showing that this type of neologicism does not live up to its philosophical promises, see, e.g., Dummett [1998] or Kolman [2008], pp. 316–320.

[12]Leśniewski's proof survived only in the account presented in Sobociński [1949]. Sobociński reports that Leśniewski arrived at his proof in 1938.

[13]The only additional assumption required for the proof of inconsistency of the amended system is that there are at least two objects. As Frege's truth values are considered to be objects, the acceptance of a universe comprising just a single object means a collapse of the object True and the object False into a single object.

[14]Cf. Kolman [2002], p. 233

[15]A name of a course-of-values will not be a logical component of a sentence in the same way in which 'car' is not a logical component of 'cart'. This illustration is taken from Potter [2000], p. 180.

construed by Frege as a special kind of courses-of-values, it follows that we lose the ability to generalize about numbers. Interestingly enough, Russell himself, in a brief note added to Appendix A to *The Principles of Mathematics* devoted to the discussion of Frege's *Grundgesetze*, expressed the opinion that this was very likely 'the true solution'.[16] An elaboration upon this general strategy leads to Zermelian axiomatic set theory.

The other strategy is to keep the abstraction principle intact but, in Frege's words, to deny the validity of the law of excluded middle for courses-of-values. It would no longer be the case that a course-of-values could be either an object or nothing but it could be a different kind of an object, i.e., an 'improper' object. This implies, in effect, giving up the unity of objects and accepting a stratification of objects into fundamentally distinct types. Frege counters that this strategy would require distinguishing a great multiplicity of different types of functions according to the levels of arguments they can take, which, in turn, would increase the complexity of the system of types of objects. Moreover, besides the excessive complexity Frege doubts that there can be some general "legislation" for deciding what are the admissible types of arguments and values for what functions.[17] Without such an underlying principle justifying the division into types, the whole solution would be hopelessly ad hoc. As suspected, this is the general direction taken by Russell in the theory of types.

The two strategies have important and profoundly different consequences for the question of what it is to be an object. They share the assumption that we can admit without difficulty ground level entities that are not associated with any functions or concepts or that are not conceived of as collections of other objects. This is the level of individuals or urelements. Traditionally, the extension of a concept is the class of all the objects that fall under this concept. Thus when we are given a concept, we are given a means for determining, for every individual object, whether it belongs to the extension of this concept or not. To grasp a concept is to be in possession of a criterion, or a *rule*. On the other hand, there is another way of specifying a collection of objects, which consists simply in determining, for every given object, whether it belongs to the collection or not. This time, we are not applying a rule, we are not classifying objects; it is the *membership* in the collection *itself* that determines the collection and that is taken as basic. We may call the collections specified by a rule 'logical', and speak of them as of 'classes', as opposed to the collection specified by listing their members which may be named 'combinatorial' and addressed as 'sets'.[18] In the remaining sections of this chapter, we will be dealing with Russell's ramified theory of types,

[16] Cf. Russell [1903], p. 522.

[17] Cf. Frege [1903], p. 255.

[18] For a more recent discussion of the distinction between classes and sets or between the logical and combinatorial collections, see Parsons [1974], Maddy [1990], pp. 102–106, and Lavine [1994], pp. 63–98.

and hence with the "logical", concept-based approach. To accentuate the difference between the two conceptions, let us say just a few words about the "combinatorial" contender.

Cantor's early set theory, as developed in his *Grundlagen einer allgemeinen Mannigfaltigkeitslehre* (Cantor [1883]), can be seen as exploiting the combinatorial notion of set. A fundamental concept applicable to sets is that of well-ordering, which Cantor defines as follows: a set a is WELL-ORDERED by a binary relation R if every strictly bounded subset of a has an immediate successor in a.[19] The concept of well-ordering is taken to be absolutely central to set theory:

> The concept of *well-ordered set* turns to be fundamental for the entire theory of manifolds. In a later article I shall discuss the law of thought that says that it is always possible to bring any *well-defined* set into the *form* of a *well-ordered* set—a law which seems to me fundamental and momentous and quite astonishing by reason of its general validity. (Cantor [1883], p. 886; Cantor's emphasis; Ewald's translation.)

This assertion is an expression of the so-called 'well-ordering principle': every set can be well-ordered.[20] Conversely: whatever cannot be well-ordered is not a set. The notion of well-ordering played a key role in Cantor's development of transfinite set theory. Interesting as it is, for our present purposes it is only essential to see its close relationship with the notion of counting. A well-ordering of a set unequivocally determines the succession of all its elements. Therefore, if we are given a well-ordered set, we can count it 'by following that well-ordering' (cf. Lavine [1994], p. 53).[21] Conversely, if we have counted all the members of a set, we have, in effect, produced its well-ordering; we have imposed a specific order on its elements. Now, given the fact that a set is determined by its members, it is essential that we have a

[19] A subset b of a is STRICTLY BOUNDED in a if there is an $x \in a$ such that x is greater than every member of b. A subset b of a has an IMMEDIATE SUCCESSOR in a if there is the least $x \in a$ such that x is greater than all the members of b.

The above is a slightly modified version of the definition in Cantor [1883], p. 884. Nowadays, it is more common to define well-ordering differently: a set a is TOTALLY ORDERED by a binary relation R if R is irreflexive and transitive, and if R satisfies 'trichotomy' on a, namely: for any $x, y \in a$, it is either the case that xRy, or yRx, or $x = y$. If a totally ordered set a also satisfies the requirement that every non-empty subset of a has a least element, it is said to be well-ordered. Cantor's original definition is essentially equivalent to the contemporary one; see Lévy [1979], pp. 38–39, for a proof.

[20] Cantor never gave a proof for his well-ordering principle, so it remained merely a fundamental assumption underlying his set theory. In 1900, Hilbert included it in his famous list of mathematical problems (as a part of the first problem, 'Cantor's problem of the cardinal number of the continuum', cf. Hilbert [1900], pp. 263–264). The proof was given four years later by Zermelo (in Zermelo [1904]) but it provoked strong reactions due to the explicit use of the axiom of choice.

[21] Cf. also Floyd and Kanamori [2006], p. 419.

way of listing or counting its members. Therefore, if there is a totality whose members cannot be counted, it cannot be regarded as a set. To conclude, for Cantor of the *Grundlagen*, whatever is a set can be well-ordered, and whatever cannot be counted is not a set.

Because Cantor's sets are conceived of as possessing this twin property of being well-orderable and subject to counting, they are not affected by the paradoxes in the same way that Frege's courses-of-values are. Let us illustrate the matter on Burali-Forti's paradox, sometimes also referred to as the 'paradox of the largest ordinal', the first of the notorious set-theoretic paradoxes.[22] It concerns the notion of ordinal number, which can be defined in the following way. Let us say that two sets have the same ORDER TYPE if they are order-isomorphic (i.e., if there is an order-preserving one-to-one correspondence between them). Then the ORDINAL NUMBER is defined as an order type of a well-ordered set.[23] Cantor proved that ordinal numbers can be totally ordered.[24] Now let Ω be the collection of all ordinal numbers and let us suppose that it is a well-ordered set. Given this assumption, Ω must have an order type, i.e., a specific ordinal number, say γ. By definition, $\gamma \in \Omega$. This means, however, that the ordinal number γ determines a subset of all the ordinals in Ω that are smaller than γ; but this subset has no other ordinal number than γ itself. Therefore, the ordinal number of Ω—which is, by definition, equal to γ—has to be, at the same time, greater than γ, which is a contradiction.

Burali-Forti himself took the paradox to show, by reductio ad absurdum, that ordinal numbers cannot be totally ordered, which contradicted the aforementioned theorem of Cantor's. What was Cantor's reaction to the paradox? In fact, he did not consider it to have any detrimental effect on his theory of sets at all. It is not that we should cease conceiving of the ordinal numbers as totally ordered; we merely need to realize that the totality of ordinal numbers is not a set. We have already seen why: precisely because it cannot be counted, i.e., it cannot be numbered by an ordinal number.[25] It lies beyond the realm of sets; the word Cantor uses in the *Grundlagen*

[22]The paradox carries the name of Cesare Burali-Forti who published an argument in which the paradox was implicit in Burali-Forti [1897], although it was apparently discovered earlier by Cantor himself. Burali-Forti stated the paradox not in terms of the relation of well-ordering, which he presented erroneously, but in terms of a more comprehensive relation of 'perfect ordering'. Despite a corrective note published in the same year, the outcome of Burali-Forti's article was rather a confused controversy than a fruitful debate (cf. Copi [1958], p. 281).

[23]This definition is based on Cantor's first 'Beiträge' (Cantor [1895], p. 497). See also Dauben [1979], p. 184.

[24]Cf. Cantor [1897], p. 216.

[25]To be more specific, in the *Grundlagen* the requirement that every set should be (in the aforementioned sense) countable takes on the form of two principles of generation, according to which every ordinal number has an immediate successor, and every set has a least upper bound (Cantor [1883], pp. 907–909). Neither of the principles of generation of numbers is applicable to the totality of ordinal numbers.

to refer to this totality is the 'absolute'.[26] Later, in the letter to Dedekind dated 3 August 1899, he famously states: 'The system Ω of all numbers is an inconsistent, absolutely infinite multiplicity' (Ewald [1996], p. 933). And not only that Ω cannot be a set, but any collection that has the same order-type as Ω is also an inconsistent multiplicity, as well as any collection into which Ω is injectable.

There is no abstraction principle at work here, so we are not forced into recognizing any object corresponding to the property of being an ordinal number. Sets are well-orderable, countable, and not just arbitrary collections. However, the question must be faced: What are sets? How are they given to us? Leaving aside the problem of how Cantor himself conceived of sets, which is a rather complex issue in itself, this question needs to be posed in the most general fashion. Having repudiated an abstraction principle that would open to us a whole domain of abstract objects, we are left with conditions and stipulations, whose development ultimately leads to axiomatic set theory. Sets simply have to be assumed to exist as the objects that satisfy the axioms we have laid down for them to satisfy, and the membership relation has to be taken as primitive. But this does not amount to anything else than to accepting set theory as a specific discipline standing on its own feet, with assumptions going far beyond those that can be subsumed under the wings of logic. This might be fair enough but, from the logicist point of view, it is an outright capitulation and an acquiescence that logicism has failed.

2.3 Russell's Diagnosis

Russell's logicist conviction, at least at around 1903, was even stronger than Frege's, who restricted his claims only to arithmetic. In *The Principles of Mathematics* Russell declares:

> All mathematics, we may say—and in proof of our assertion we have the actual development of the subject—is deducible from the primitive propositions of formal logic: these being admitted, no further assumptions are required. (Russell [1903], §434, p. 458.)

Of course, the proof consisting in the 'actual development of the subject' was something with which Russell struggled for the whole following decade. After trying virtually every possible response to the paradoxes, and after developing his theory of denoting, he ended up with the ramified theory of types. Fortunately, we will not deal with these intermediate struggles; we

[26] Thus Burali-Forti's paradox can be seen to have, in a peculiar sense, a positive effect on Cantor's set theory. As Kanamori puts it, Cantor 'used it *positively* to give mathematical expression to his Absolute' (Kanamori [1996], p. 13).

will merely outline Russell's response to the paradoxes and the rationale justifying his conception of the hierarchy of propositional functions.

Russell believed that the individual paradoxes were not disparate strokes of bad luck. Rather, they should be viewed as closely related symptoms of a single disease that inevitably afflicts certain attempts to build mathematics on a logical basis. In other words, the paradoxes reveal that there is something fundamentally wrong with the conception of logical objects. The furthest point of Russell's effort to uncover a common root of the paradoxes is reached in the article 'On Some Difficulties in the Theory of Transfinite Numbers and Order Types' (Russell [1906*b*]). There (pp. 142–143) he comes up with the following schema for a property φ and a function f:[27]

$$\forall u\big(\forall x[x \in u \to \varphi(x)] \to [\exists z(z = f(u)) \wedge \varphi(f(u)) \wedge f(u) \notin u]\big). \quad \text{(S-Com)}$$

In English: for any class u, if all members of u are φ, then there exists an object z assigned to the class u by the function f such that z is φ but it is not a member of u. This, as it stands, is no contradiction. Assuming that all u's members are φ, (S-Com) merely states that not everything that is φ belongs to u, and that the value that f assigns to u is a witness.

What is remarkable about the function f of this schema is that it can be used to generate a well-ordered sequence φ whose ordering corresponds to that of Cantor's series of ordinal numbers. How can this be done? Assume a function f and an x such that f assigns a value to x, i.e., such that $\exists z(z = f(x))$. Let us take $f(x)$ to be the only member of u. It follows that we have defined the first term of the sequence φ, and we have obtained a class $u = \{f(x)\}$. Incidentally, we have confirmed that the antecedent of the implication is satisfied, which means that the right-hand side of (S-Com) holds, i.e., $\exists z(z = f(u)) \wedge f(u) \notin u$. This gives us the next term of the sequence φ, namely $f(u)$, so our sequence φ has already two members: $f(x)$ and $f(\{f(x)\})$. Put them both into a class, and apply the right-hand condition again to obtain the third term $f(\{f(x), f(\{f(x)\})\})$, etc. In this way we obtain a sequence of immediate successors. If u is a class of elements of which none is the greatest, the next term in the sequence is the least upper bound of u. This represents Cantor's two principles of generation, and produces a sequence 'ordinally similar to that of all ordinals' (Russell [1906*b*], p. 143).

We have said that (S-Com), as it stands, represents no contradiction. However, if it is combined with the additional assumption that there is a class $\{x \mid \varphi(x)\}$ which is in the range of $\forall u$, the contradiction is immediate. Burali-Forti's paradox results if $\varphi(x)$ is 'x is an ordinal number', and $f(x)$ is the least ordinal number greater than every ordinal in x. Russell's paradox results if $\varphi(x)$ is '$x \notin x$', and $f(x) = x$. Note that without the assumption

[27] Russell [1906*b*] did not present anything of this symbolically. I build upon the symbolic rendering that can be found in Kanamori [1997], p. 295.

that any φ determines a class, we do not obtain Russell's paradox but merely a sequence of classes that do not belong to themselves, without there being any single class collecting them all. Russell does not provide examples of how other paradoxes can be put into this common form, and we will not attempt to answer the question whether they all really share the suggested structure. We will merely add two more examples. First, Berry's paradox of 'the smallest natural number not denoted by any expression of English of fewer than seventeen words'. The paradox can be obtained from (S-Com) if we take $\varphi(x)$ to be 'x is a numerically determinate expression of English', and $f(x)$ to be the expression 'the smallest natural number not denoted by any member of x of fewer than seventeen words'. Secondly, an analogue of the liar paradox for the concept of arithmetical truth. Let us take $\varphi(x)$ to be 'x is the Gödel number of a truth of arithmetic', and $f(x)$ to be the Gödel number (say, g) of the diagonalization of the formula $\neg\psi(x)$, where $\psi(x)$ defines an arithmetic set of Gödel numbers of truths of arithmetic. Then g is the Gödel number of a truth of arithmetic but is not a member of the set defined by $\psi(x)$.

The force of the schema (S-Com) comes from the fact that it represents a key principle standing behind the whole theory of ordinal numbers and, consequently, a core component of set theory as a mathematical discipline. Furthermore, it appears perfectly unobjectionable as a tool for defining certain complex properties or sequences of objects. Nonetheless, it is incompatible with the twin requirement that every property should determine a class, and that the range of the individual variable should be unrestricted. It is worth at this point to quote Russell in full.

> [T]he contradictions result from the fact that, according to current logical assumptions, there are what we may call *self-reproductive* processes and classes. That is, there are some properties such that, given any class of terms all having such a property, we can always define a new term also having the property in question. Hence we can never collect *all* the terms having the said property into a whole; because, whenever we hope we have them all, the collection which we have immediately proceeds to generate a new term also having the said property. (Russell [1906*b*], p. 144; Russell's emphasis.)

The conclusion is that not every property determines a class. However, we have seen that this already became obvious in connection with Frege's Basic Law V. The main virtue of Russell's analysis is that it attempts to formulate an underlying principle that will make it possible to uniformly distinguish between the concepts that do determine classes and the "rogue" ones that do not, i.e., those that do not have determinate extensions. Russell's diagnosis states that without determinate extensions are those concepts that are self-reproductive. We could also employ Dummett's term 'indefinitely

extensible', which is better known.[28] A concept φ is considered to be self-reproductive if there is a function f such that the condition (S-Com) holds. Note, however, that it does not necessarily follow that self-reproductive concepts do not possess extensions. It only follows that the extension of a self-reproductive concept cannot be a class that itself falls under this very concept. Seen from this perspective, the trouble with Basic Law V does *not* consist in positing that every function has a course-of-values or that every concept has an extension.[29] It rather lies in assuming that every course-of-value and every extension can be construed as a class, i.e., as a well-determined object falling within the range of the individual variable. Hence, according to Russell's analysis, Frege's system ended in ruins because it failed to recognize the existence and the extraordinary nature of self-reproductive concepts.

The starting point of Russell's logicist solution to the paradoxes is thus the distinction between the concepts or, to be more faithful to Russell's terminology, propositional functions of one variable that do determine a class, for which Russell introduces the term 'predicative',[30] and those that do not, i.e., those that are impredicative. The latter are, indeed, the self-reproductive propositional functions. As we have seen, self-reproductive concepts may be conceived of as generating an unbounded sequence of classes, and if they are taken only as such, they do not lead to a contradiction, not at least in any straightforward way. However, the contradiction is immediate once we attempt to encompass the whole sequence and reapply the given concept to it. It is this attempt to break the sequence by a circle that brings about the collapse.

Although a form of circularity is clearly implicit in Russell's analysis, the idea that the unifying cause of the individual contradictions lies in a vicious circle comes from Poincaré. In 1905-1906 he published the first two parts of his three-part series entitled 'Les mathématiques et la logique' (Poincaré [1905] and Poincaré [1906*a*]). Then he read Russell [1906*b*] (and a letter by Zermelo replying to his two articles), and in response published the third sequel bearing the same name (Poincaré [1906*b*]), developing (in section IX)

[28] The notion of indefinite extensibility has been a persistent feature of Dummett's writings ever since Dummett [1963]. In Dummett [1993*b*], p. 441, an indefinitely extensible concept is characterized as 'one such that, if we can form a definite conception of a totality all of whose members fall under that concept, we can, by reference to that totality, characterize a larger totality all of whose members fall under it.'

[29] This does not entail, however, that everything can be rectified by a simple modification of Basic Law V such as: $\grave{\epsilon}F(\epsilon) = \grave{\alpha}G(\alpha) \leftrightarrow (Self\text{-}reprod.(F) \wedge Self\text{-}reprod.(G)) \vee \forall x(F(x) \leftrightarrow G(x)))$. As Shapiro and Wright [2006], pp. 284–285, point out, it is not clear how to characterize self-reproductiveness (or indefinite extensibility) in purely logical terms. Moreover, such a modified version of Basic Law V would, taken on its own, fail to provide for infinite objects.

[30] Russell [1906*b*], p. 146, states: 'We define a *predicative* propositional function as one which determines a class (or a relation, if it contains two variables).' This is the first of three different meanings of the word 'predicative' we encounter in this chapter.

the notion of a vicious circle in connection with the paradoxes of Richard's and Burali-Forti's. (Besides, Poincaré found viciously circular also Cantor's definition of $\aleph_1$.) The definitions (i.e., not concepts) that commit a vicious circle are called 'non-predicative'.[31] However, apart from identifying the vicious circularity itself, Poincaré's analysis of the phenomenon has little more to offer to a logicist. Nonetheless, in his reply to Poincaré called 'Les paradoxes de la logique' (Russell [1906*a*]), Russell concedes to the vicious-circle diagnosis, and formulates a so-called 'vicious-circle principle' (op. cit., p. 198).

Unfortunately, we are facing here a rather problematic and much debated issue concerning how exactly the vicious-circle principle should be understood. Not only Russell does not manage to articulate the principle unambiguously but he states it somewhat differently at different places. In Russell [1908] alone we can find at least four distinct versions (on pp. 75, 101, and two on p. 63).[32] Anyway, instead of striving to decipher the true meaning of the vicious-circle principle and reconstruct the intended theory behind it, we will take an easier route. We will sketch a particular interpretation, as coherent as it stands, of the principle together with the ontology underlying Russell's conception of logic. Still, merely outlines of a broader picture are drawn in this section. Technical aspects of the ramified theory of types are postponed to section 2.4.

Let us take as paradigmatic the two following formulations of the vicious-circle principle that appear in *Principia* (Russell and Whitehead [1962], p. 39 and p. 40, respectively; my emphasis):

> [G]iven any set of objects such that, if we suppose the set to have a total, it will contain members which *presuppose* this total, then such a set cannot have a total. (VCP-P)
>
> If, provided a certain collection had a total, it would have members only *definable* in terms of that total, then the said collection has no total. (VCP-D)

We will read the words 'a collection has a total' as meaning that the given collection may be taken as a single object, that the expression designating it is subject to the substitution rules, and that it may be quantified over.[33] To put it simply, a propositional function φ has a total if it determines a class.

Now how are we supposed to understand what (VCP-P) and (VCP-D) demand? Let us first render (VCP-P). It has been rather convincingly argued by Jung [1999], pp. 60–61, that presupposition in (VCP-P) is the

[31] This is the second meaning of the word 'predicative' that appears in this chapter.

[32] Some have claimed that Russell, strictly speaking, does not formulate a single vicious-circle principle but a number of them. Gödel, for instance, uncovers three (Gödel [1944], p. 135).

[33] Here we follow Jung [1999], p. 60.

relation of ontological dependence which can be further characterized as being asymmetric, transitive, and having the strength of supervenience. Jung suggest that we should construe the requirement imposed by (VCP-P) as equivalent to well-foundedness in set theory,[34] and paraphrases it in this way: 'all totalities that may be treated as objects are well-founded' (p. 61). In particular, (VCP-P) prohibits chains of the following kind: $f_1(f_2)$, $f_2(f_3)$,..., $f_{n-1}(f_n)$, and $f_n(f_1)$.

The latter formulation (VCP-D) is slightly more complex. First, it contains the term 'definable'. As Quine points out, it is hardly definability that is in question here since a definition primarily concerns notation; the circularity of (VCP-D) thus must not be understood as smuggling the definiendum into the definiens.[35] It makes more sense to think of what really goes on here in terms of a specification of classes combined with the assumption of class existence, which is nothing else than the application of an abstraction principle. We will say that to SPECIFY a propositional function f is to provide a formula of the form: $\exists f \forall x_1, \ldots, x_n (f(x_1, \ldots, x_n) \leftrightarrow \varphi)$ (where φ does not contain 'f' free), and we will read (VCP-D) as meaning specifiability instead of definability. Secondly, it is clear from other statements by Russell that 'in terms of' needs to be understood as 'by means of quantification over'. Thus (VCP-D) commands: no totality t may contain an object that is specifiable only by means of quantification over t.[36]

Of the two formulations, (VCP-P) certainly appears more plausible. An analogue of this principle has been adopted in standard systems of set theory, and, taken on its own, (VCP-P) provides justification for a simple theory of types. On the other hand, (VCP-D), justifying the ramification of the type theory, has been subject to strong criticism. The main charge has been that (VCP-D) may be acceptable only if we understand our ontology constructivistically, i.e., if, by specifying an object, we are somehow creating it. Once we assume that objects exist independently of our specifications—viz. Ramsey's example of the tallest in a group or Quine's most typical Yale man[37]—there does not seem to be any reason for prohibiting specifications quantifying over all objects. Certainly, Russell was no constructivist, at least not the Russell of the period between the *Principles* and *Principia*. How can he hold (VCP-D) then?

In his remarkable paper, Jung [1999], pp. 67–74, identifies two logico-ontological assumptions that, if adopted, justify (VCP-D) on the realist reading. The first assumption is that the logical form of the formula specifying a proposition or a propositional function as an object corresponds to the ontological form of the object being specified in the sense that to the logical components of the formula there correspond ontological components of

[34] The concept of well-foundedness in set theory is treated in section 3.1.

[35] Quine [1969], pp. 242–243.

[36] Here again I am indebted to the analysis and arguments in Jung [1999], pp. 65–67.

[37] Cf. Ramsey [1925], p. 204, and Quine [1969], p. 243, respectively.

the object. Among other consequences, this assumption permits a stronger formulation of the principle of abstraction suitable for intensional contexts. Instead of the extensional equivalence one is entitled to the intensional use of identity: $\exists f(f = \varphi(\hat{x}_1, \ldots, \hat{x}_n))$ (where φ does not contain 'f' free). Consequently, as identity of propositions and propositional functions is no longer extensional, our assumption yields intensional abstraction principles, which make it possible to distinguish between two propositions with the same truth value or two propositional functions satisfiable under the same conditions. The other assumption identified by Jung is the assertion that a variable presupposes the terms that make up its range.[38] For instance, the propositional function $\forall x(f(x, \hat{y}))$ presupposes the range of the variable x, and the range of x presupposes its members. Now the two logico-ontological assumptions just described reveal the logical relationship between (VCP-P) and (VCP-D). For it is easy to see that (VCP-P) plus the two assumptions implies (VCP-D): assume that (VCP-D) is not true. Then there will be a formula specifying a function f which will contain a quantified variable, say g, whose range contains f. By the first principle, g will be a genuine component of f, hence f presupposes g. Yet, by the second principle, g presupposes f. These two results put together contradict our construal of (VCP-P).[39] Therefore, if we accept (VCP-P) as well as the two principles identified by Jung, we have to accept also (VCP-D). Of the two renderings of the vicious-circle principle, (VCP-P) is the one that is more general, and (VCP-D)—once supplemented by the two assumptions—is merely a special case of (VCP-P).

It is time to sum up. When searching for a common root of the paradoxes, Russell comes to realize that the contradiction brought about by Frege's Basic Law V is caused by our failure to recognize the existence and peculiar nature of self-reproductive concepts. These concepts themselves neither involve nor give rise to immediate contradiction, and there does not seem to be any firm ground for denying that they have an extension. What does lead to the contradiction is taking the extensions of these concepts to be objects over which our variables range. It is this additional assumption which, therefore, must be banned. And the ban takes the form of the vicious-circle principle.

Let us move on and describe in some detail the outcome of Russell's attempt to produce a system of logic that would succumb to the vicious-circle principle, in both of its versions (or, which is the same, in its more general version supplemented by the two aforementioned assumptions), namely his ramified theory of types.

[38] A similar claim, presented as a "speculation", can be found in Goldfarb [1989], p. 37: 'Russell takes a variable to presuppose the full extent of its range.'

[39] Cf. Jung [1999], p. 72.

2.4 The Ramified Theory of Types

It is well known that Russell's progress towards the ramified theory of types was anything but straightforward. First, in *The Principles of Mathematics*, he came up with a "tentative" version of the simple theory of types, which, however, he himself showed to be inconsistent, the culprit being nothing else than the paradox of propositions.[40] Three years after the publication of the *Principles*, in the article 'On Some Difficulties in the Theory of Transfinite Numbers and Order Types', he presented three other solutions to the paradoxes, namely the "zigzag" theory, the theory of the limitation of size and the substitutional theory.[41] Nonetheless, none of the suggested solutions was found acceptable. Eventually, Russell turned again to the theory of types but this time the types were no longer simple. The result appeared as 'Mathematical Logic as Based on the Theory of Types' (Russell [1908]).

A coherent presentation of Russell's theory of types is made rather troublesome by several facts. Firstly, in the years to come Russell put forward other, slightly different versions of the theory of types in the article 'La théorie des types logiques' (Russell [1910]) and in *Principia Mathematica* (Russell and Whitehead [1962]). To add to the confusion, the theory of types is described twice in *Principia*: first in chapter II of the introduction, which is more or less a transformation of the aforementioned French article, and then in section $*12$.[42] The fact that the two presentations of the theory of types differ in certain respects is openly acknowledged in the preface to *Principia* where it is said that the hierarchy described in $*12$ is "stricter" than the one in the introduction, and that it 'is that which is assumed throughout the rest of the book'.[43] Secondly, Russell, relying on the concept of 'typical ambiguity', never produced a symbolism for specifying individual types and orders. Modern treatments of the notation for the ramified theory of types have thus to be inventive. Perhaps the most commonly used is the one due to Church [1976], though there are other formulations.[44] Thirdly, Russell's treatment of the foundations of logic is 'greatly lacking in formal precision' (Gödel [1944], p. 126). Extraction of precise contents from sometimes quite confusing statements may often be a rather difficult task.

Fortunately, as we are interested only in a particular aspect of Russell's theory of types, we will not take positions on the various interpretative issues the theory poses. Our goal is just to formulate a coherent, broadly Russellian

[40] Cf. Russell [1903], pp. 523–528. This paradox is discussed in section 2.5.

[41] Cf. Russell [1906*b*], pp. 144–156.

[42] Cf. Russell and Whitehead [1962], pp. 37–65 (chapter II of the introduction) and pp. 161–167 (section $*12$).

[43] See Russell and Whitehead [1962], p. vii. For a discussion of the differences between the accounts of the theory of types, cf. Linsky [1999], pp. 73–88.

[44] Cf., for example, Hatcher [1968], Copi [1971], Chihara [1973], pp. 19–23, Myhill [1979] or Kamareddine *et al.* [2002], pp. 199–230. A concise comparison of the former four particular formulations with that of Church's can be found in Linsky [1999], pp. 66–72.

theory, and not to engage in historical investigations or in the consideration of remote technical aspects. In what follows, we will outline a version of the ramified theory of types based on the formulation presented in Church [1976], pp. 747–751. A major difference is that whereas Church assigns the ramified types to variables, i.e., expressions, we will primarily attribute them to entities over which the variables range, and only secondarily to the variables themselves.

The key idea behind the ramified types—or 'r-types', as Church calls them to distinguish them from the types of the simple theory of types—is that entities are not stratified into different types only on the basis of the character of the entities that can fill in the empty argument places (which would yield the simple types) but also on the basis of the character of all the other entities that the given entity presupposes in a broader sense (this leads to the notion of order). Unsurprisingly, the r-types are thus determined in a rather complex way. The R-TYPE of an entity is defined inductively as follows:

- There is an r-type ι for individuals.
- Suppose that $m \geq 0$, $n \geq 1$ and $\beta_1, \ldots, \beta_m$ are any r-types; then there is an r-type $(\beta_1, \beta_2, \ldots, \beta_m)/n$ to which belong m-ary propositional functions of level n with the arguments of r-types $\beta_1, \ldots, \beta_m$, respectively.

The ORDER of an entity is just a natural number. It is defined inductively in the following way:

- The order of an individual is 0.
- The order of a propositional function of r-type $(\beta_1, \beta_2, \ldots, \beta_m)/n$ is $k + n$ where k is the greatest of the orders corresponding to the r-types $\beta_1, \ldots, \beta_m$, and if $m = 0$, $k = 0$.

The value of the auxiliary notion of LEVEL n, occurring after the slash in the designations of all r-types except those for individuals, is determined as follows. Assume a propositional function of r-type $(\beta_1, \beta_2, \ldots, \beta_m)/n$, and let k be the greatest of the orders of $\beta_1, \ldots, \beta_m$ and j the greatest of the orders of the bound variables occurring in the definition of this propositional function. Then $n = 1$ if $j \leq k$, and $n = j + 1$ if $j > k$.

According to the theory of types developed by Russell, the ramified types are mutually exclusive: whatever belongs to one r-type does not belong to any other.[45] In other words, a propositional function containing an empty

[45] In Russell and Whitehead [1962], p. 161, we read:

> In virtue of ∗9·14, if φx, φy, and ψx are significant, *i.e.* either true or false, so is ψy. From this it follows that two types which have a common member coincide, and that two different types are mutually exclusive.

There is evidence that Russell considered the possibility of construing the hierarchy as

argument space that requires an argument of a certain ramified type is considered applicable only to arguments of that given type but not to arguments of lower types. Contrary to this decision of Russell's, Church constructs his system of r-types as cumulative: a given r-type is to serve only as a ceiling of the range of significance of a propositional function; it is not to coincide with it. Cumulativeness is brought in the system by means of the relation of being DIRECTLY LOWER: the r-type $(\alpha_1, \alpha_2, \ldots, \alpha_m)/k$ is directly lower than the r-type $(\beta_1, \beta_2, \ldots, \beta_m)/n$ if $\alpha_1 = \beta_1, \alpha_2 = \beta_2, \ldots, \alpha_m = \beta_m$, and if $k < n$. (Note that the relation of being directly lower does not translate into the simple comparison of orders in the sense that the r-type of propositional functions of order $< n$ would be directly lower than that of propositional functions of order n. For instance, the second-order r-type $((\iota)/1)/1$ is directly lower that the third-order r-type $((\iota)/1)/2$, but it is not directly lower than the third-order r-types $(\iota)/3$ and $(((\iota)/1)/1)/1$.) Once we have the relation of being directly lower, we obtain cumulativeness by stating that if a propositional function can take as arguments members of a certain r-type, it can also take as arguments members of any directly lower r-type.

The notation we have introduced allows for assigning r-types not only to propositional functions, whose number of empty argument places is at least one, but also to propositions, which do not contain any empty argument places at all. The r-type of propositions is thus $()/n$, where n is the appropriate level. It is the level indicator that prevents the collapse of all propositions into a single type, and which gives rise to a subhierarchy of propositions within the hierarchy of r-types.[46]

A terminological remark. A propositional function that does not contain any bound ("apparent") variables but only free ("real") variables is said to be a MATRIX. A propositional function whose level indicator n equals to 1 is called PREDICATIVE.[47] From these specifications it follows that any matrix

cumulative (cf. Potter [2000], pp. 146–147) but rejected it without providing any compelling argument. The rationale Russell gave in a letter to Hawtrey, a friend of his who recommended to him to take the path of cumulativeness, consisted in nothing more than 'a sort of symbolic instinct, which I rely upon more than I can explicitly justify' (quoted from Potter [2000], p. 146).

Let us also point out that the decision between the cumulative and non-cumulative construction of the hierarchy has some interesting consequences. Peressini [1997] shows on the examples of Grelling's and Bouleus paradoxes that the choice to understand the r-types cumulatively allows us to formulate certain questions—which may or may not be answered with the resources at our disposal—that are not formulable in the non-cumulative theory. Still more importantly, it can be shown that the non-cumulative syntax makes problematic some class symbols that are perfectly acceptable if types are construed cumulatively. For details, see Peressini [1997], pp. 394–396.

[46]The decision to accept the hierarchy of propositions is somewhat controversial. See note 52 on p. 32.

[47]This is a new, third and final meaning of the word 'predicative', related to the two meanings mentioned in section 2.3. In what follows, we will use this word in the sense defined here.

as well as any first-order propositional function is predicative.[48]

In order to visualize how the orders, levels and r-types all fit together, let us represent their relationship graphically in the following table:[49]

orders	r-types
...	...
5	$(\iota)/5$ ...
4	$(\iota)/4$ $((\iota)/1)/3$ $((\iota)/2)/2$ $(((\iota)/1)/1)/2$ $((\iota)/3)/1$ $((\iota)/1)/2)/1$ $(((\iota)/2)/1)/1$ $((((\iota)/1)/1)/1)/1$
3	$(\iota)/3$ $((\iota)/1)/2$ $((\iota)/2)/1$ $(((\iota)/1)/1)/1$
2	$(\iota)/2$ $((\iota)/1)/1$
1	$(\iota)/1$
0	ι

The table clearly pictures that the number of different r-types of each order grows exponentially: in general, for $n > 0$ there are 2^{n-1} r-types of each order n.

There are, in general, three basic ways in which propositional functions of higher r-types can be obtained from the function of lower r-types. First, simply by putting functions of higher orders together in a matrix, such as $f(g(\hat{z}), x)$, which has the r-type $((\iota)/1, \iota)/1$. Secondly, by quantifying (in the definition introducing the given function) over predicative functions. For example, $\exists g(f(g(\hat{z}), x))$ of r-type $(\iota)/2$ is obtained by prefixing a quantifier to our predicative matrix ranging over the predicative functions $g(\hat{z})$. (We are using the word 'matrix' here to denote the expression, which makes the word ambiguous. Yet, no confusion should arise as the intended meaning should always be clear from the context.) The third method is based on abstraction.[50]

[48]It needs to be noted that it is not always clear how the expressions 'matrix' and 'predicative function' are meant to be used in *Principia*. Both are introduced in the Introduction, cf. Russell and Whitehead [1962], p. 50 and p. 53, respectively. However, in section ∗12, we read that 'a function is said to be *predicative* when it is a matrix' (p. 164), which goes against what we said above. The whole issue is discussed by Linsky [1999], pp. 77–85, who reaches the conclusion that the restriction of the intended meaning of the expression 'predicative function' is to do with the strictness of the theory of types of ∗12, which is then rebalanced by the greater strength of the axiom of reducibility.

[49]The table is based on the table in Chihara [1973], p. 21, Chihara's notation having been replaced by Church's.

[50]In fact, Russell assumed that propositional functions of higher types are obtained exclusively by the first two of the three ways mentioned. In particular, he claimed that any non-predicative function is obtained from a predicative matrix by quantifying over some of the variables in the argument positions ranging over the predicative functions (cf. Russell and Whitehead [1962], p. 54). There has been a whole discussion about whether this is a result of a mere oversight on the part of Russell or whether warding off certain functions was a deliberate decision.

To see how abstraction gives rise to a propositional function which cannot be obtained by quantifying over predicative functions, consider the example mentioned in Hylton

So far we have introduced the system of r-types as a hierarchy of propositional functions, sticking out upwards from the ground level of individuals. That is to say, we have dealt mainly with the ontological aspect of the theory of types. It remains to outline the main features of the logical dimension of the theory of types, i.e., it remains to describe the language and the deductive apparatus as based on the particular construal of the type theory given above. We will not attempt to reconstruct the actual system devised in *Principia Mathematica*, though, nor will we describe the system in full. A standard system of many-sorted predicate calculus with identity will be supposed, and only peculiar features related to the adoption of the ramified theory of types will be explicitly elaborated.

The language of the ramified theory of types will be designated as '$\mathcal{L}_{\mathrm{RTT}}$'. It is assumed that it contains the usual quantifiers and propositional connectives, and that there is an unlimited supply of variables of each r-type. The formation rules of $\mathcal{L}_{\mathrm{RTT}}$ require that $f(x_1, \ldots, x_m)$ is a well-formed formula only if the r-type of f is $(\beta_1, \ldots, \beta_m)/n$ and the r-types of the variables $x_1, \ldots, x_m$ are $\beta_1, \ldots, \beta_m$, respectively, or any respective directly lower r-types. The circumflex notation is used to form abstracts for propositional functions: '$f(\hat{x}_1, \ldots, \hat{x}_m)$' is a term, and '$f(\hat{x}_1, \ldots, \hat{x}_m)(y_1, \ldots, y_m)$' is a well-formed formula. Quantification over propositions is permitted, and a propositional variable standing alone is considered to be a well-formed formula.

The deductive system contains—besides the standard axioms and rules of inference—three additional axiom schemata of abstraction.[51] The first is an abstraction principle for propositional functions:

$$\exists f \forall x_1, \ldots, x_m (f(x_1, \ldots, x_m) \leftrightarrow \varphi), \qquad \text{(A-CoF)}$$

where f is a functional variable of r-type $(\beta_1, \ldots, \beta_m)/n$, $x_1, \ldots, x_m$ are variables of r-types $\beta_1, \ldots, \beta_m$, respectively, and the order of the bound variables occurring in φ is less than that of f, the order of the free variables and constants occurring in φ is less than or equal to that of f, and f does not occur free in φ. The second axiom schema is an abstraction principle for propositions:

$$\exists p (p \leftrightarrow \varphi), \qquad \text{(A-CoP)}$$

[1990], p. 309 (attributed to Thomas G. Ricketts). Take the statements: $\forall x^{\iota}(R^{(\iota)/1}(x) \rightarrow P^{(\iota)/1}(x)) \wedge P(a)$ and $\forall x^{\iota}(R^{(\iota)/1}(x) \rightarrow Q^{(\iota)/1}(x)) \wedge Q(b)$, where '$P$', '$Q$' and '$R$' are predicate constants and 'a' and 'b' are individual constants. By existential generalization, we get the following statements: $\exists f^{(\iota)/1}[\forall x^{\iota}(R^{(\iota)/1}(x) \rightarrow f(x)) \wedge f(a)]$ and $\exists f^{(\iota)/1}[\forall x^{\iota}(r^{(\iota)/1}(x) \rightarrow f(x)) \wedge f(b)]$. Then, by abstracting from the individuals a and b, we obtain the propositional function $\exists f^{(\iota)/1}[\forall x^{\iota}(r^{(\iota)/1}(x) \rightarrow f(x)) \wedge f(\hat{z})]$ whose r-type is $(\iota)/2$, which holds of both a and b. Hence we may conclude that $\exists g^{(\iota)/2}(g(a) \wedge g(b))$. This propositional function $g(\hat{z})$ of r-type $(\iota)/2$ has not been obtained by quantification over predicative functions.

[51] The first two schemata are presented in Church [1976], p. 750. The remaining one is based on $*9{\cdot}15$, cf. Myhill [1979], p. 82, or Linsky [1999], p. 60.

where p is a propositional variable of r-type $()/n$, the order of all the bound variables occurring in φ is less than n, and the order of the constants occurring in φ is less than or equal to n.[52] The remaining schema is the abstraction principle generating terms that may be subject to existential generalization:

$$\forall y_1, \ldots, y_m(\varphi(\hat{x}_1, \ldots, \hat{x}_m)(y_1, \ldots, y_m) \leftrightarrow \varphi), \qquad \text{(A-CoT)}$$

where φ is a propositional function of r-type $(\beta_1, \ldots, \beta_m)/n$, and $y_1, \ldots, y_m$ are variables of r-types $\beta_1, \ldots, \beta_m$, respectively.

As the axiom schemata (A-CoF), (A-CoP) and (A-CoT) employ the connective for material equivalence, '$\leftrightarrow$', they are obviously extensional. This is to say that they guarantee the existence of a propositional function that is coextensive or of a proposition that is just extensionally equivalent to a given formula. If we wanted to express a stronger, intensional relation, we would need to introduce a symbol for equality of higher-order entities, that is, either we could extend the permissible use of the symbol '$=$' to propositional functions and propositions, or we could follow Church, and supply a new symbol such as '$\equiv$'. The extensional axiom schemata (A-CoF), (A-CoP) and (A-CoT) could then be transformed into the intensional ones by replacing the biconditional '$\leftrightarrow$' by the chosen symbol. The decision taken up here is to extend the use of the symbol '$=$' to propositions but to keep the schemata extensional, as the intensional versions are not required for our specific purposes.

To obtain a full system of the ramified theory of types which would be adequate for rendering classical mathematics, it is also necessary to suppose some form of the axiom of choice, and add the notorious axioms of infin-

[52] It is true that the introduction of the abstraction principle for propositions appears to go against the spirit of *Principia* in which quantification over propositions is not used apart from very few exceptions such as in $*14{\cdot}3$; and there we can read that the quantification over propositions is not legitimate 'without the explicit introduction of the hierarchy of propositions with a reducibility axiom' (Russell and Whitehead [1962], p. 185). The suspicion that propositions might not be seen in *Principia* as fully-fledged, autonomous entities on a par with individuals and propositional functions is further strengthened by the statement that the symbols for propositions are considered 'incomplete symbols' (op. cit., p. 44). There are some, e.g., Cocchiarella [1989], p. 45, who consequently refuse to accept propositions as genuine entities, and repudiate them from the hierarchy of r-types. However, from the fact that symbols for propositions are incomplete it does not follow that they are 'contextually defined' or that they disappear on analysis, i.e., that we can get rid of assuming propositions as entities. (For an illuminating discussion of these three interrelated notions see Neale [2001], pp. 224–232.) The fact of the matter is that Russell's logical reconstruction of mathematics based on the contextual analysis of class terms requires only the hierarchy of propositional functions; the hierarchy of propositions is not needed. For this reason, as Linsky [1999], p. 58, puts it, the decision not to quantify over propositions 'seems a matter of convenience' rather than 'some eliminativist ontological position'. Therefore, we follow Church, and we keep propositions among the genuine entities.

ity and reducibility.[53] We will designate the full system including all the aforementioned axioms as 'RTT'.

2.5 Ramified Types and Propositions

The purpose of this section is twofold: firstly, to illustrate the functioning of the ramified theory of types—in particular its treatment of propositions—on a sample paradox, and secondly, to face a widespread criticism of the need for ramification of types first formulated by Ramsey. Nonetheless, the present section may be viewed merely as an appendix to the previous section which contained the substantial development.

The example that will be presented is Russell's paradox of propositions. It is discussed in §500 of *The Principles of Mathematics* (Russell [1903], p. 527) but it is not mentioned in the list of paradoxes in Russell [1908], pp. 59–64, or in Russell and Whitehead [1962], pp. 60–65, and it apparently does not reappear in Russell's subsequent writings.[54] The paradox involves a construction analogical to the one used in Russell's paradox, and it leads to the conclusion that there cannot be a class of all propositions.[55] The paradox presupposes that there are such entities as propositions and that these entities are intensional.[56] The criterion of identity for propositions is to be the identity of their constituents:

$$\forall f(f(\varphi) = f(\psi)) \rightarrow \varphi = \psi. \qquad \text{(S-IdP)}$$

This is nothing else than a form of the principle of identity of indiscernibles, the more controversial half of so-called 'Leibniz's law'. However, once we assume that there are such entities as propositions, and that these entities are intensional, (S-IdP) follows.

We are ready for the paradox.[57] Let us define a propositional function

[53]The axiom of infinity is stated in *Principia* in ∗120·03, the axiom of reducibility in ∗12·1 and ∗12·11. The latter is discussed below in section 2.7.

[54]Cf. Goldfarb [1989], p. 39.

[55]The paradox is sometimes called 'Russell-Myhill Paradox' owing to the fact that, half a century later, it was independently rediscovered by John Myhill (cf. Myhill [1958]) who found it to affect Church's intensional system developed in Church [1951].

[56]Note that it does not matter whether we use the word 'propositions'. What the paradox requires is the existence of intensional entities that are referred to by sentences, be it facts or whatever else. This paradox thus does not directly affect Frege's system since, for Frege, sentences were names of the truth values, not of propositions. In the correspondence with Frege, Russell attempted to reformulate the paradox so that it affected Frege's theory of senses ('Sinne') but, in the end, he accepted that this cannot be done (cf. Landini [1992], p. 168). For a modern attempt at a formulation of this paradox in the way that would show the inconsistency of Frege's theory of senses, see Klement [2001], pp. 18–24.

[57]In this reconstruction of the paradox, I build more or less on the formal treatment found in Linsky [1999], pp. 63–66, and Potter [2000], pp. 131–134, who do without classes. For a version of this paradox with classes, which is perhaps slightly closer to Russell's original formulation, see Landini [1992], pp. 163–168, or Urquhart [2003], pp. 288–289.

$f(\hat{\varphi})$:

$$f(\varphi) =_{Def.} \forall p(\varphi(p) \rightarrow p). \qquad \text{(D-PPr}_1\text{)}$$

$f(\hat{\varphi})$ is a function that takes as an argument a propositional function $\varphi(\hat{p})$ such that every proposition that is φ is true. To put the same thing differently, $f(\hat{\varphi})$ assigns a value (i.e., a proposition) to a propositional function of propositions, i.e., $f(\hat{\varphi})$ is a function from functions of propositions to propositions. This is, indeed, a familiar situation: $f(\hat{\varphi})$ is just another attempt to establish, in effect, a one-to-one correspondence between propositions and classes of propositions. Inevitably, given the assumption (S-IdP), a contradiction is bound to be round the corner.

To see how it can be derived, consider the propositional function $w(\hat{\varphi})$, defined as follows:

$$w(\varphi) =_{Def.} \exists m(\varphi = f(m) \wedge \neg m(\varphi)). \qquad \text{(D-PPr}_2\text{)}$$

Now take the proposition $f(w)$ and ask whether $f(w)$ is w. Assume first that $\neg w(f(w))$. Then—by simply putting '$f(w)$' for φ in (D-PPr$_2$) and by applying the negation to the right-hand side—we obtain:

$$\neg w(f(w)) \leftrightarrow \forall m[f(w) = f(m) \rightarrow m(f(w))].$$

As the propositional function $w(\hat{\varphi})$ lies within the range of $\forall m$, we can get $f(w) = f(w) \rightarrow w(f(w))$, which is the same as $w(f(w))$. The assumption that $f(w)$ is not w thus leads to a contradiction. So assume the contrary, i.e., that $w(f(w))$. Then this holds:

$$w(f(w)) \leftrightarrow \exists m[f(w) = f(m) \wedge \neg m(f(w))].$$

From this, however, together with the principle (S-IdP), it immediately follows that $\neg w(f(w))$. Again, a contradiction.

How is the paradox blocked in the the ramified theory of types? Consider what the r-type of $f(\hat{\varphi})$ is. Given its definition (D-PPr$_1$), it must be at least $((()/1)/1)/1$, i.e., third-order. Now let us add the r-type superscripts to the definition (D-PPr$_2$):

$$w^{(()/1)/2}(\varphi) \leftrightarrow \exists m^{(()/1)/1}(\varphi = f^{((()/1)/1)/1}(m) \wedge \neg m(\varphi)).$$

Assuming that φ is a first-order proposition, $w(\hat{\varphi})$ is a third-order propositional function taking first-order propositions as arguments. However, as $f(\hat{\varphi})$ is, in this case, also a third-order function, it follows that these two functions cannot be combined. If we wanted to assert that $f(w)$, we would need another propositional function $f^*((\hat{\varphi}))$ at least of r-type $(((()/1)/2)/1$, i.e., a fourth-order function. Similarly, we would need a new, sixth-order function $w^*(\hat{\varphi})$ to obtain $w^*(f^*(w))$. Hence the expressions '$f(w)$' and '$w(f(w))$' are not well-formed, and the derivation of the paradox is no longer possible.

The paradox of propositions is noteworthy because, on the assumption that there are intensional entities denoted by sentences, it represents a counterexample to Ramsey's influential division of paradoxes into the set-theoretic paradoxes on the one hand and the semantic paradoxes on the other.[58] The former affect set theory (or the theory of classes), and can be easily solved by a simple theory of types, i.e., a type theory without the level indicators. The paradoxes of the second group, according to Ramsey, 'cannot be stated in logical terms alone; for they all contain some reference to thought, language, or symbolism, which are not formal but empirical terms' (Ramsey [1925], p. 183). Ramsey's conclusion is, to put it bluntly, that we do not need to care about the latter paradoxes since they belong to the subject matter of special, empirical sciences.

The paradox of propositions does not make use of any extra-logical vocabulary. Propositions are meant to be logical objects, which is to say that they are to be the subject matter of logic, and not of any special science. The paradox would thus presumably fall among the logical (and, in its alternative formulation for classes, set-theoretic) paradoxes. Yet, it is not solved by the simple theory of types. To see this, it is sufficient to realize that, according to the simple theory of types, all propositions belong to the simple type (). Therefore, the proposition $f(w)$ is a legitimate argument for the propositional function $w^{(())}$, so we can legitimately assert that $w(f(w))$. Then, however, there is nothing that can prevent the reconstruction of the paradox. To block the paradox, we need, in addition to the hierarchy of propositional functions, a hierarchy of propositions. Yet the whole point of the simple theory of types is to remove this additional hierarchy.

We may conclude that, unless we take the assumption that sentences denote some sort of intensional entities to be a claim of a special or empirical discipline or unless we refute it as ill-founded, Ramsey's claim that the simple theory of types is sufficient for solving the logical paradoxes has to be regarded as incorrect.

2.6 Truth in the Ramified Theory of Types

The system of the ramified theory of types as described above is able to make assertions about propositional functions designated by the function terms of $\mathcal{L}_{\mathrm{RTT}}$ and propositions expressed by the sentences of $\mathcal{L}_{\mathrm{RTT}}$. Obviously, since it does not contain names of its own expressions and sentences, it cannot directly express anything about its own syntax. However, RTT is a

[58] Cf. Ramsey [1925], pp. 183–184, 187–192. To be more precise, Ramsey calls the first group of paradoxes 'logical or mathematical' and claims that, without proper measures taken to prevent their appearance, 'they would occur in a logical or mathematical system itself' (op. cit., p. 183). However, since that time set theory and logic have become clearly separated, and as most of the paradoxes of the first group involve set-theoretic notions, it is now customary to call them 'set-theoretic' rather than 'logical'.

rather powerful theory; it allows for the development of arithmetic, analysis or large parts of set theory. Consequently, it is sufficiently rich for the introduction of a coding scheme that would associate natural numbers or other objects it can speak of with the syntactic objects of $\mathcal{L}_{\mathrm{RTT}}$ such as symbols and formulas. In other words, its syntax can be arithmetized. We will not, however, follow the path of arithmetization in this section, and we will not develop any such coding. A reason for this decision is that this path is followed throughout chapter 4, a system of arithmetization being described in section 4.1. Moreover, it is interesting in its own right to see how the semantic notions may be brought into the system of RTT without the technique of arithmetization. It is just important to remember that arithmetization can be carried out in $\mathcal{L}_{\mathrm{RTT}}$, so it is not necessary to introduce any special primitive constants, which we are going to do in the subsequent paragraphs.

Yet before anything else, we will make one more assumption concerning the syntactic objects of $\mathcal{L}_{\mathrm{RTT}}$, namely we will suppose that our variables of r-type ι range also over symbols and formulas of $\mathcal{L}_{\mathrm{RTT}}$, i.e., that syntactic objects are among the individuals. The rationale behind this decision is that we do not wish to impose any unnecessary restrictions on the domain of individuals, and as symbols and formulas are well defined objects, there is no reason to exclude them. Thus the variables of r-type ι range, among other things, over formulas and sentences of $\mathcal{L}_{\mathrm{RTT}}$, while the variables of appropriate higher-order r-types range over the entities designated or expressed by these syntactic objects, namely propositional functions and propositions.[59] So, in $\mathcal{L}_{\mathrm{RTT}}$, we can speak about the symbols and expressions as well as about the designated or expressed entities. Is there a way to exploit this capacity and throw some light on the very expressive powers of language? Well, yes. But there is still something missing. What is missing is a craftily forged link connecting these different kinds of entities. This link will now be introduced by means of the primitive relation constant 'Val', accompanied by meaning postulates governing its use.[60]

To be precise, a special relation Val needs to be introduced for each r-type (of order greater than 1). With the superscript indicating the given

[59]Later, in his article 'Logical Atomism', Russell himself explicitly endorsed the view that all syntactic objects, in contrast to their meanings, belong to a single r-type: 'All words are of the same logical type; a word is a class of series, of noises or shapes according as it is heard or read. But the meanings of words are of various different types' (Russell [1924], p. 332). Russell does not state that the single type to which all words belong is the type of individuals but this choice seems decisively the most natural.

[60]The relation constant Val in its full generality is introduced in Church [1976], pp. 754–756. There are, indeed, other ways to go about this "missing link", such as by means of a designation relation constant introduced by Myhill [1979], p. 86, or an 'expresses' constant discussed in Hazen [1983], pp. 371–373. However, our aims are broader than those of Myhill's and Hazen's, which makes the constant Val more suitable.

r-type, Val becomes:

$$Val^{(\iota,\ldots,\iota,(\beta_1,\ldots,\beta_m)/n)/1},$$

where $m \geq 0$, $n \geq 1$, $\beta_1, \ldots, \beta_m$ are any r-types, and the number of ι's preceding '$(\beta_1, \ldots, \beta_m)/n$' is $m+1$. The r-type indicator reveals that the relation Val holds between $m+1$ individuals and a propositional function of r-type $(\beta_1, \ldots, \beta_m)/n$. If $m = 0$, it holds between a single individual and a proposition $()/n$. Note that Val is predicative, i.e., its level is always 1. If k is the order of the propositional function of r-type $(\beta_1, \ldots, \beta_m)/n$, the order of Val is $k+1$. In some cases, it will suffice to indicate only the order of Val by means of a superscript instead of its full r-type. This yields the sequence: $Val^2, Val^3, \ldots, Val^{k+1}$.

What is the primitive constant Val supposed to mean? Let f be a propositional function of r-type $(\beta_1, \ldots, \beta_m)/n$, let $a^\iota_1, \ldots, a^\iota_m$ range over the variables of r-types $\beta_1, \ldots, \beta_m$, respectively, and let the value of v^ι be a well-formed formula with only the variables $a^\iota_1, \ldots, a^\iota_m$ free. Observe that $a^\iota_1, \ldots, a^\iota_m, v^\iota$ all range over variables taken as symbols, i.e., as individuals of r-type ι. Given these assumptions, the following well-formed formula of $\mathcal{L}_{\mathrm{RTT}}$:

$$Val^{(\iota,\ldots,\iota,(\beta_1,\ldots,\beta_m)/n)/1}(a^\iota_1, \ldots, a^\iota_m, v^\iota, f^{(\beta_1,\ldots,\beta_m)/n}) \qquad \text{(P-Val)}$$

is intended to mean that for every assignment of the values $x^{\beta_1}_1, \ldots, x^{\beta_m}_m$ to the variables $a^\iota_1, \ldots, a^\iota_m$ occurring in the formula v^ι as free, the value of v^ι is $f^{(\beta_1,\ldots,\beta_m)/n}(x^{\beta_1}_1, \ldots, x^{\beta_m}_m)$. To put it simply, disregarding the r-type indicators, the relation Val relates a formula and its free variables (i.e., syntactic objects treated as individuals) with the appropriate entities from the hierarchy of r-types that are assigned to them as their values. So Val is supposed to provide the aforementioned missing link between the syntactic objects and their values.

The meaning of Val that we have just specified informally can be fixed formally by means of three special axioms, or 'meaning postulates'.[61] To make the statements of the axioms shorter and more easily readable, we will always assume that the r-type of f is $(\beta_1, \ldots, \beta_m)/n$, that the r-type of g is $(\beta_1, \ldots, \beta_m)/k$, and that the r-type of Val is the lowest possible, namely $(\iota, \ldots, \iota, (\beta_1, \ldots, \beta_m)/n \text{ or } k)/1$. The first axiom is the principle of univocacy of Val:

$$\forall a_1, \ldots, a_m \forall v \forall f \big(Val(a_1, \ldots, a_m, v, f) \rightarrow \forall g [Val(a_1, \ldots, a_m, v, g) \rightarrow \forall x_1, \ldots, x_m (f(x_1, \ldots, x_m) \leftrightarrow g(x_1, \ldots, x_m))] \big). \qquad (\text{A-Val}_1)$$

In simple English, the same formulas and free variables are assigned the same values. This axiom, as it asserts the equivalence of two propositional functions, is extensional but it can be turned into a stronger, intensional principle

[61] Again, we are following Church [1976], p. 755.

which asserts the identity of the two functions. The intensional formulation is obtained by replacing '$\forall x_1, \ldots, x_m(f(x_1, \ldots, x_m) \leftrightarrow g(x_1, \ldots, x_m))$' with '$f = g$'. In what follows, we will only need the weaker, extensional version of the axiom.

The second axiom is the following comprehension schema:

$$\exists a_1, \ldots, a_m \exists v \exists f \big(Val(a_1, \ldots, a_m, v, f) \wedge \forall x_1, \ldots, x_m [f(x_1, \ldots, x_m) \leftrightarrow \varphi] \big), \qquad \text{(A-Val}_2\text{)}$$

where φ is a formula containing only the variables $x_1, \ldots, x_m$ free, whose bound variables are all of order less than that of f, and whose constants are all of order less than or equal to that of f. The schema (A-Val$_2$) says that any formula φ satisfying these restrictions is equivalent to some formula $f(x_1, \ldots, x_m)$ and that there are syntactic objects $a_1, \ldots, a_m, v$ such that $Val(a_1, \ldots, a_m, v, f)$. A stronger, intensional version of this axiom is obtained if the biconditional '$\leftrightarrow$' gets replaced with identity '$=$'. As with the axiom (A-Val$_1$), we will not need the intensional version of (A-Val$_2$).

Finally, the remaining axiom:

$$Val^{n+1}(a_1, \ldots, a_m, v, f) \rightarrow Val^{k}(a_1, \ldots, a_m, v, f), \qquad \text{(A-Val}_3\text{)}$$

where $k > n + 1$, expresses the cumulativeness of the relation Val. Now it is true that we have said that Val is to be predicative. However, Val's being predicative just signifies that its level number is 1. If, for instance, Val^k is introduced for $f^{(\iota)/2}$, its r-type will be $(\iota, \iota, (\iota)/2)/1$ of order 3. Yet, owing to the cumulativeness property introduced on p. 29, $Val^{(\iota,\iota,(\iota)/2)/1}$ will also be applicable to $f^{(\iota)/1}$, which is directly lower than $f^{(\iota)/2}$, and to which we would primarily apply $Val^{(\iota,\iota,(\iota)/1)/1}$ of order 2. The goal of the axiom (A-Val$_3$) is then nothing else than to explicitly express the requirement that the predicative relation Val should be employed cumulatively. If we consider cumulativeness to be a fully general feature of the formation rules of $\mathcal{L}_{\mathrm{RTT}}$, the axiom (A-Val$_3$) is superfluous, and may be dropped.

Let us call the ramified theory of types with the primitive constant Val and the axioms (A-Val$_1$), (A-Val$_2$), and (A-Val$_3$) 'RTT+Val'. Having the primitive relation Val in hand, we are immediately able to define the relations of satisfaction and truth.[62] Let us start with satisfaction, which can be defined as follows:

$$Sat^{(\iota,\ldots,\iota,\beta_1,\ldots,\beta_m,\iota)/n+1}(a_1^{\iota}, \ldots, a_m^{\iota}, x_1^{\beta_1}, \ldots, x_m^{\beta_m}, v^{\iota}) \leftrightarrow_{Def.}$$
$$\exists f^{(\beta_1,\ldots,\beta_m)/n}(Val^{(\iota,\ldots,\iota,(\beta_1,\ldots,\beta_m)/n)/1}(a_1, \ldots, a_m, v, f) \wedge f(x_1, \ldots, x_m)). \qquad \text{(D-Sat}_{\mathrm{r}}\text{)}$$

[62] The definitions of Sat and Tr that follow are taken from Church [1976], pp. 756–757, with the difference that Church presents them only with the implication connective, i.e., merely as expressing a necessary condition.

In English, (D-Sat$_\mathrm{r}$) says that a formula v containing the variables $a_1, \ldots, a_m$ free is satisfied by the entities $x_1, \ldots, x_m$ if and only there is a propositional function f such that $Val(a_1, \ldots, a_m, v, f)$, and f holds of $x_1, \ldots, x_m$. It is important to understand how the notion of order works in this definition. Let the order of f be $k+n$, where k is the greatest of the orders of $x_1, \ldots, x_m$. Then the order of Val is $k+n+1$. Now, despite the fact that the arguments of Sat are only of r-types $\iota, \ldots, \iota, \beta_1, \ldots, \beta_m, \iota$, the order of Sat is not just $k+1$. The reason is that Sat is an abbreviation for its defining formula, which contains quantification over f whose order depends partly on its bound variables, indicated by n. Hence the order of Sat has to be the same as that of Val, namely $k + n + 1$.

Although it does not differ from (D-Sat$_\mathrm{r}$) in substance, the definition of truth looks much neater:

$$Tr^{(\iota)/n+1}(v^\iota) \leftrightarrow_{Def.} \exists p^{()/n}(Val^{(\iota,()/n)/1}(v, p) \wedge p). \qquad \text{(D-Tr}_\mathrm{r}\text{)}$$

I.e., a sentence v is true if and only if it expresses a proposition p and it is the case that p. Here again the order of Tr equals that of Val since its definition includes quantification over p.

Two closely related aspects of these definitions deserve a particular attention. Firstly, the property Tr is not a property of propositions but a property of sentences, understood as syntactic objects, i.e., individuals of r-type ι. Thus there is no discrimination between sentences with respect to the property of truth: Tr of any order can be *meaningfully* applied to any sentence whatsoever. If it is applied to a sentence expressing a proposition of an incorrect order, the result is not a meaningless sentence but a false sentence. For instance, let v be a true sentence expressing a second-order proposition. The second-order property Tr^2 is, indeed, applicable to v but the defining condition $\exists p^1(Val^2(v, p) \wedge p)$ is not fulfilled since it is only the case that $\exists p^2(Val^3(v, p))$. The sentence '$Tr^2(v)$' is thus false, and it fails to express any proposition at all.[63] Similarly for the relation Sat. The relation $Sat(a_1, \ldots, a_m, x_1, \ldots, x_m, v)$ is a relation that holds between a formula

[63] In spite of the fact that there is nothing, as far as I know, in *Principia Mathematica* and other writings concerned with the ramified theory of types that would suggest that Russell might have intended a definition of truth along the lines proposed by Church, we seem to be in agreement with Russell's general remarks on truth in the section III of chapter II of the Introduction to *Principia* (Russell and Whitehead [1962], pp. 41–47). There we can read:

> [W]hen we judge 'a has the relation R to b,' our judgment is said to be *true* when there is a complex 'a-in-the-relation-R-to-b,' and is said to be *false* when this is not the case. This is a definition of truth and falsehood in relation to judgments of this kind. (Russell and Whitehead's emphasis; p. 43.)

Here, if we construe the 'complex' as 'proposition', and if we understand the relation Val as explicitly furnishing the relation of 'expresses' which is present only implicitly in the structural similarity of the sentence 'a has the relation R to b' to the proposition 'a-in-the-relation-R-to-b', our definition (D-Tr$_\mathrm{r}$) becomes virtually a verbatim transcript of the

and its free variables on the one hand—which are all syntactic objects belonging to the r-type ι—and appropriate entities of different r-types on the other. Assume that v expresses that a second-order propositional function $f^{(\iota)/2}$ holds. Then the statement that $Sat^{(\iota,\iota,\iota)/2}(a^\iota, x^\iota, v^\iota)$ is meaningful but it is false as it is an abbreviation for $\exists f^1(Val^2(a, v, f) \wedge f(x^0))$. Owing to cumulativeness, however, both Tr and Sat of higher orders can hold of sentences or formulas expressing propositions or propositional functions of much lower orders. For instance, $Tr^5(v^\iota)$ is true of a sentence v expressing a true first-order proposition.

It is this intricate relationship between sentences and formulas, i.e., syntactic objects, and entities of other appropriate r-types, mediated by the constant Val, that makes the introduction of concepts such as truth or satisfaction an interesting and significant accomplishment. The relations Val, Sat or Tr are valuable precisely as tools for investigating the relationship between the language taken as a system of symbols and a system of entities designated or expressed by the symbols or their combinations. Were we interested only in the truth itself, i.e., in the actual true propositions without any regard to sentences expressing them, we would not need to introduce any explicit truth predicate at all. The system of RTT as we described it in section 2.4 makes it possible to assert that a proposition is true without employing the property of truth at all. The fact that a proposition having a property φ is true can be easily expressed as:

$$\varphi(p) \rightarrow p. \qquad \text{(S-ImT)}$$

Similarly, saying that a proposition with a property φ is false is rendered just as: $\varphi(p) \rightarrow \neg p$. Indeed, this concept of truth, which is implicit in the logic of *Principia Mathematica*, is something very different from our Tr, which is a property of sentences taken as objects of r-type ι. It is presumably the latter that can help investigate the semantics of language.

The second aspect deserving attention concerns the multitude of the relations Val, Sat and Tr. A property Tr^n can be defined for any order $n \geq 2$, and as we take r-types to be cumulative, each property Tr^n can also hold of all the sentences expressing propositions of orders $< n$. Yet, as there is no greatest order, there is no single property Tr whose extension would include all true sentences whatsoever. Similarly for satisfaction. So there is a sequence of the distinct relations Sat and Tr, each of a different order and each holding only of formulas or sentences expressing propositional functions or propositions of lower orders, but there is no single relation Sat or Tr expressible in $\mathcal{L}_{\text{RTT+Val}}$ or definable in RTT+Val which would be true of formulas or sentences expressing propositional functions or propositions of any order whatsoever. Our definitions have thus fulfilled Russell's dictum that 'the words "true" and "false" have many different meanings, according

quoted characterization of truth.

to the kind of proposition to which they are applied' (Russell and Whitehead [1962], p. 42).[64]

Thus there is no way we can establish a single class of all true sentences of $\mathcal{L}_{\text{RTT+Val}}$; however, having the properties $Tr^2, Tr^3, \ldots, Tr^i, \ldots$, we can, in a rather straightforward manner, establish such a class for all the sentences expressing propositions of a given order. To do this, though, we need to assume the principle of bivalence for sentences to the effect that any well-formed sentence is either true or false. If we grant this assumption, it is easy to set up a classification of sentences of $\mathcal{L}_{\text{RTT+Val}}$, taken as objects of r-type ι, according to the order of propositions they aim to express. Employing the constant Val, we may introduce the following relation: $Oer^{n+1}(v^\iota, n) =_{Def.} \exists p^n(Val^{n+1}(v, p) \vee Val(v, \neg p))$, which assigns to each sentence v the order n of the proposition it aims to express. With this relation in hand, we can specify a class $\{v \mid Oer(v, 1)\}$ of all the sentences that express first-order propositions or their negations; similarly for second-order propositions, etc. In this fashion, we obtain a comprehensive classification of sentences (i.e., individuals of order 0) according to the orders of propositions they aim to express. On the basis of this classification, it is easy to define the subclass of all true sentences belonging to any given class. For the sentences expressing propositions of order 1 (or their negations) it is $\{v \mid Oer(v, 1) \wedge Tr^2(v)\}$, and similarly for any higher order. To conclude, we are able to determine a sequence of classes comprising all sentences that express propositions of a given order (or their negations) as well as a sequence of classes containing all true sentences of a given class, and—which is the most important matter to realize—we are able to do all this *within* one and the same language $\mathcal{L}_{\text{RTT+Val}}$.

To be sure, there is nothing that prevents us from making a decision to separate our single language $\mathcal{L}_{\text{RTT+Val}}$ into a hierarchy of languages of increasing orders. This is the path taken up by Church [1976], who permits the language of order n to contain only bound variables of order $n-1$ and free variables of order n. Then a language L_{n+1} of order $n+1$—which contains the full language L_n plus the variables ranging over the r-types of higher-order objects that are not allowed in L_n and the appropriate Val constants—can be considered to be a 'semantical meta-language of L_n' (op. cit., p. 756). Nevertheless, from what we described in the preceding paragraphs, it follows that the talk of different languages is not essential; it is merely a result of a decision to give to one's investigations a format that is deemed convenient. Church is explicit about this:

> [I]t is quite indifferent whether we speak of a single language L

[64] Of course, the property Tr has been defined as applicable to sentences, and not to propositions; therefore, we would have to read this statement as: 'according to the kind of proposition expressed by the sentence to which they are applied'. However, the quotation well supports our replacement of a single concept of truth with a multitude of different concepts, which I take to be the more substantial part of its message.

> and a hierarchy of orders of variables and predicates within it or whether we speak of an infinite hierarchy of languages L_1, L_2, L_3, ..., as it is evident that the distinction is merely terminological. (Church [1976], p. 756.)

2.7 Reducibility and Expressibility

A remarkable feature of Russell's conception of the hierarchy of propositional functions is revealed by analyzing the effects of the axiom of reducibility. This ill-famed axiom, stated in ∗12·1 and ∗12·11 of *Principia*, can be rendered as follows:

$$\exists f^{(\beta_1,\ldots,\beta_n)/1}\forall x_1,\ldots,x_n(\varphi(x_1,\ldots,x_n)\leftrightarrow f(x_1,\ldots,x_n)). \qquad \text{(A-Red)}$$

It asserts that for any propositional function $\varphi(\hat{x}_1,\ldots,\hat{x}_n)$ there is an extensionally equivalent predicative propositional function $f(\hat{x}_1,\ldots,\hat{x}_n)$, i.e., one whose r-type is the lowest compatible with the r-types of the arguments. On the logicist reading, (A-Red) is a powerful abstraction principle asserting the existence of entities without specifying them. (On the other hand, if we assume the existence of a full hierarchy of propositional functions beforehand, and approach the axiom of reducibility from such a non-logicist point of view, its role becomes similar to that of a comprehension schema of set theory such as the axiom of separation of ZFC.[65] But obviously, Russell needs the strong, logicist reading of the axiom.)

The axiom of reducibility is required to make the system of *Principia* adequate for classical mathematics. To understand why, let us cite Russell's eliminative definition of classes (∗20·01):

$$f\{z \mid \psi(z)\} =_{Def.} \exists\varphi(\forall x[\varphi!(x)\leftrightarrow\psi(x)]\wedge f(\varphi!(\hat{z}))). \qquad \text{(D-Cla)}$$

We have used '!', Russell's "typically ambiguous" symbol for a predicative function, rather than our r-type indicators to make (D-Cla) more compact. It says that a class of ψs is f just in case there is a predicative function φ coextensive with ψ which is f. In brief, classes are identified with the predicative functions coextensive with their defining properties. (To obtain a full eliminative theory of classes, we would also want a definition of the membership relation $\in$ and definitions permitting the elimination of bound class variables. These all appear in ∗20·02, ∗20·07 and ∗20·071 of *Principia*; but we will not discuss them here.) In the light of (D-Cla) it is clear that, in contrast with propositional functions, the talk of classes is extensional.

The definition of classes is the core of Russell's "no-class" theory, an attempt to do without classes as objects. According to this theory, 'class' is merely a locution belonging to a mode of speech that can always be

[65] This was observed by Gödel [1944], pp. 140–141.

eliminated, and transformed back into the "real" talk of propositional functions. Now, just in this paragraph, let us remain within the bounds of this "virtual" talk of classes—it makes the deficiency of the system of *Principia* without reducibility quite easy to articulate. In this virtual talk, the introduction of classes amounts to the introduction of a whole hierarchy of virtual objects *within* the hierarchy of propositional functions. Hence within the broader, more robust hierarchy of intensionally construed propositional functions there can be identified a thinner, virtual subhierarchy of extensionally construed entities. Once we have got the subhierarchy of classes, we may investigate whether it is adequate for conveying classical mathematics. Assume that standard set theory is adequate for mathematics. Can the same be said also of this theory of virtual classes? The answer is: No, unless we add the axiom of reducibility. The reason why the theory of virtual classes cannot stand on its own lies in the fact that it does not supply enough classes; the virtual subhierarchy does not contain some classes that are required. The most quoted example of a central mathematical notion that cannot be obtained without reducibility is the least upper bound of a bounded class of real numbers. The failure consists in the fact that if we attempt to specify the least upper bound without reducibility, we always end up with an impredicative propositional function and, as a result, with an entity of a higher order than that of the members of the class. A predicative propositional function would get us all that is needed but we do not seem to be able to specify any. So we have to turn to the axiom of reducibility for help. Indeed, the whole purpose of this axiom is precisely to supply all the predicative functions we need, so that the lacunae in the hierarchy of classes are filled.

There is no pretense that the axiom of reducibility is satisfactory. The problem that affects it has two sides. Firstly, how can an assertion of the existence of all these predicative functions be justified? Note that owing to the abstraction axiom schemata (A-CoF) and (A-CoT) the hierarchy of functions already includes every single propositional function we are able to pick out by a specifying formula. The very point of the axiom of reducibility is to enlarge the amount of functions available by positing also functions for which we do not possess a specification. This is a notorious difficulty, one that Russell himself was well aware of and that significantly contributed to the eventual demise of logicism. Yet, as this problem is well known, we will not discuss it here any further. Instead, we will deal with the other side of the coin, namely with the issue of expressibility. Provided that the axiom (A-Red) guarantees the existence of all possible predicative functions, even those for which we do not have specifications, the conclusion seems inevitable that our language, $\mathcal{L}_{\mathrm{RTT}}$ (possibly with added constants for all specifiable functions), fails to accommodate them all, i.e., to provide a means for designating them by constants of appropriate r-types or picking them out by specifying formulas. Evidently, they can be quantified over, so we can

express that they exist but can we express what exactly they are? After all, if we were forced to admit entities that cannot be designated and about which not enough facts are expressible, what would be left of the logicist project? Would not we get something bordering on the metaphysics of entities beyond all specification, rather than a logical analysis of what is given to us in language? It is this concern that is dealt with in the rest of this section.

There is a paradox which has particularly significant consequences that throw light on the issue of expressibility of predicative propositional functions whose existence is guaranteed by RTT, namely Grelling's "heterological" paradox.[66] Moreover, a careful analysis of this paradox helps to refute the charge that the axiom of reducibility reinstates some of the paradoxes that are blocked in the system without reducibility.[67] We will build upon the formalization presented in Church [1976], pp. 751–754, 758–760. Let us start by the definition of the following constants:

$$Het^{n+1}(v) =_{Def.} \exists a \exists f^{(\iota)/n}(Val^{n+1}(a, v, f) \wedge \neg f(v)), \tag{D-Het}$$

where all the variables without r-type or level indicators are of r-type ι. (D-Het) says that, disregarding the orders, a predicate v containing one free variable is Het if and only if it designates a propositional function $f(\hat{x})$ but itself is not f. In other words, a predicate (a syntactic object) is Het if it does not have the property it expresses.

We will formalize neither the paradox nor the solution to it in RTT+Val in full. We will merely list the main three theorems that can be obtained without the axiom of reducibility. Complete proofs are supplied in Church [1976], pp. 753–754. The theorems are:

$$[Val^{m+2}(a, v, f^{(\iota)/m+1}) \wedge \forall x(f(x) \leftrightarrow Het^{m+1}(x))] \rightarrow \neg Het^{n+1}(v), \text{ if } m \geq n, \tag{S-Het$_1$}$$

$$[Val^{m+2}(a, v, f^{(\iota)/m+1}) \wedge \forall x(f(x) \leftrightarrow Het^{m+1}(x))] \rightarrow Het^{n+1}(v), \text{ if } m < n, \tag{S-Het$_2$}$$

$$\exists a, v \exists f^{(\iota)/m+1}[Val^{m+2}(a, v, f) \wedge \forall x(f(x) \leftrightarrow Het^{m+1}(x))]. \tag{S-Het$_3$}$$

Theorems (S-Het$_1$) and (S-Het$_2$) state, for each m of the associated relation Val^{m+2}, whether a predicate v does or does not have the particular property Het^{n+1}. Theorem (S-Het$_3$) asserts that there exist a variable a, a

[66] This paradox first appeared in Grelling and Nelson [1907/08] but the authorship is ascribed to Grelling.

[67] This objection was articulated by Chwistek [1921] against Richard's paradox. Against Grelling's antinomy, it was reformulated by Copi [1950]. A reply to this line of argument against the axiom of reducibility is implicit in Church [1976] and explicit in Myhill [1979] and Hazen [1983], pp. 365–375.

predicate v and an appropriate function $f(\hat{x})$, so the antecedents of the conditionals of the first two theorems are satisfied. Consequently, both of the consequents hold. There is nothing paradoxical about this situation since the level requirements make them all compatible. Obviously, a contradiction immediately arises once we remove the level indicators.

What happens when we make recourse to the axiom of reducibility? Can the paradox be reinstated, as it has been charged? Given the definition (D-Het), the axiom of reducibility yields the following existential claim:[68]

$$\exists h^{(\iota)/1} \forall v (h(v) \leftrightarrow \exists a \exists f^{(\iota)/n} [Val^{n+1}(a, v, f) \wedge \neg f(v)]), \qquad \text{(S-h}_1\text{)}$$

i.e., there is a predicative, namely first-order propositional function that is extensionally equivalent to the impredicative function Het whose order is $n + 1$. This may be put also as:

$$\exists h^{(\iota)/1} (h(x) \leftrightarrow Het^{n+1}(x)). \qquad \text{(S-h}_2\text{)}$$

We will now take '$h^{(\iota)/1}$' to be a schematic letter which can be replaced by a constant designating any propositional function satisfying statements (S-h$_1$) and (S-h$_2$), or quantified to become a regular variable. Granted this convention, we may assert that $\forall v (h^{(\iota)/1}(v) \leftrightarrow Het^{n+1}(v))$. Yet, it is then possible to derive the following analogues of theorems (S-Het$_1$) and (S-Het$_2$):

$$[Val^2(a, v, f^{(\iota)/1}) \wedge \forall x (f(x) \leftrightarrow h^{(\iota)/1}(x))] \rightarrow \neg h(v), \qquad \text{(S-h}_3\text{)}$$

$$[Val^2(a, v, f^{(\iota)/1}) \wedge \forall x (f(x) \leftrightarrow h^{(\iota)/1}(x))] \rightarrow h(v). \qquad \text{(S-h}_4\text{)}$$

However, an analogue of (S-Het$_3$) has not been derived and it does not seem derivable. Therefore, we may take the antecedent of the two implications to be false, and avoid the conclusion that $\neg h(v) \wedge h(v)$. It follows that the paradox has not been reinstated. However, we are forced to regard the antecedents to be false, and this has important consequences.

The fact of the matter is that the following result can be obtained by a simple transformation from (S-h$_3$) and (S-h$_4$): $\forall x (f(x) \leftrightarrow h^{(\iota)/1}(x)) \rightarrow \neg Val^2(a, v, f)$. By a substitution for 'h' and standard logic it is possible to derive the following theorem:

$$\forall a, v \forall h^{(\iota)/1} (\forall x [h(x) \leftrightarrow Het^{n+1}(x)] \rightarrow \neg Val^2(a, v, f)). \qquad \text{(S-h}_5\text{)}$$

In English, if $h^{(\iota)/1}(\hat{x})$ is a propositional function coextensive with Het^{n+1}, there is no predicate v (with a variable a free) such that its value is $h^{(\iota)/1}(x)$. Since, according to the result established as (S-h$_2$), such an $h^{(\iota)/1}(\hat{x})$ does exist, we have to conclude that the function $h^{(\iota)/1}(\hat{x})$ cannot be designated by any predicate v of $\mathcal{L}_{\text{RTT+Val}}$. Therefore, Russell's ramified theory of types with the axiom of reducibility, i.e., full RTT+Val, entails the existence of

[68] The argument that follows is again based on the one in Church [1976], pp. 758–760.

unnameable propositional functions. So not only are propositional functions extra-linguistic entities but there are some among them that are, in a sense, beyond the reach of language, $\mathcal{L}_{\text{RTT+Val}}$. Hence our suspicion expressed several pages back has been fulfilled.

What should we make of this curious situation? One conclusion is clear: the axiom of reducibility does not reintroduce a contradiction where RTT+Val without reducibility blocks it, at least not in connection with Grelling's paradox and not in any straightforward fashion. However, as we have just remarked, the contradiction is avoided only if it is accepted that there exist unnameable propositional functions. Still, one should be careful not to get carried away, and consider a bit more cautiously what the fact that some propositional functions are unnameable means. It obviously does not signify that nothing can be expressed about these functions. After all, we have already established several non-trivial properties of an unnameable function $h^{(\iota)/1}(\hat{x})$. The key points to realize are, firstly, that the talk of such functions in $\mathcal{L}_{\text{RTT+Val}}$ can only be extensional, and secondly, that it is is strictly hierarchical. We will very briefly comment on both these points, starting with the former.

The axiom of reducibility asserts the existence of a coextensive propositional function, and that is simply all we get. Since propositional functions are intensional, i.e., there can be different functions that assign the same values to the same arguments, any extensional talk of propositional functions obviously fails to represent intensional differences between functions and pick out functions as single entities. So we can make all sorts of different assertions about our function $h^{(\iota)/1}(\hat{x})$ but they will be valid of any other function coextensive with it.

The other point is closely connected with the first. Recall what form the extensional statements concerning $h^{(\iota)/1}(\hat{x})$ had. To take the simpler of the two biconditionals (S-h_1) and (S-h_2) stated above, $h^{(\iota)/1}(\hat{x})$ was introduced by: $\exists h^{(\iota)/1}(h(x) \leftrightarrow Het^{n+1}(x))$. The lowest order Het can have is 2, while the order of h is 1. Thus the lowest possible combination of orders we can get is the following: $\exists h^1(h(x) \leftrightarrow Het^2(x))$. It follows that the very existence of a first-order propositional function $h^1(\hat{x})$ extensionally equivalent to the given function $Het^2(\hat{x})$ is expressible by means of a second-order sentence. And given the fact that $h^1(\hat{x})$ is unnameable, we can speak of it *only* using a sentence of at least second-order. Now if we introduce another predicative (first-order) function as extensionally equivalent to impredicative $Het^3(\hat{x})$, this new unnameable function will only enter possible talk through the medium of third-order sentences. That is, second-order sentences that conveyed the introduction of $h^1(\hat{x})$ do not permit introduction of first-order functions equivalent to impredicative functions Het of orders greater than 2. So, although the individual predicative equivalents are all first-order, the talk of them is structured strictly hierarchically, alongside with that of the impredicative constants Het.

Indeed, the existence of unnameable functions should not have come to us as a surprise. As Church puts it, this fact 'not only is intelligible but even is to be expected in the light of Tarski's theorem about truth' (Church [1976], p. 759). Recall our definition (D-$\mathrm{Tr_r}$) in section 2.6. Evidently, the axiom of reducibility is applicable to the properties Tr of appropriate orders in the same way as it were to Het, so we get: $\exists t^{(\iota)/1}(t(x) \leftrightarrow Tr^{n+1}(x))$. This gives us an unnameable predicative propositional function $t^{(\iota)/1}(\hat{x})$, and the theorem (S-$\mathrm{h_5}$), as Church notes, becomes an expression of Tarski's theorem:

$$\forall a, v \forall t^{(\iota)/1}(\forall x[t(x) \leftrightarrow Tr^{n+1}(x)] \rightarrow \neg Val^2(a, v, f)).$$

So what we have established that holds of $h^{(\iota)/1}(\hat{x})$ and $Het^{n+1}(\hat{x})$, holds also of $t^{(\iota)/1}(\hat{x})$ and $Tr^{n+1}(\hat{x})$, and presumably of other semantic notions.

The issue of expressibility can also be articulated in terms of classes. A class of φs contains all objects that have the property φ. We have seen that, according to Russell's "no-class" theory of classes, in order to determine a class, it is required that we are in possession of a predicative propositional function. From the conclusions we draw above concerning the notions of Het and Tr it follows that these properties themselves, being impredicative, do not directly determine classes of predicates or classes of sentences, respectively, but they do so only via the intermediary of coextensive predicative functions such as $h^{(\iota)/1}(\hat{x})$ and $t^{(\iota)/1}(\hat{x})$. Hence, in order to establish a class of first-order truths of $\mathcal{L}_{\mathrm{RTT+Val}}$, we need to introduce a predicative function $t^1(\hat{x})$ via the impredicative second-order function $Tr^2(\hat{x})$. To obtain a class of second-order truths of $\mathcal{L}_{\mathrm{RTT+Val}}$, we need an impredicative $Tr^3(\hat{x})$, etc. In general, a class of truths of order n can only be specified using sentences of order at least $n + 1$. Similarly for Het and presumably other semantic notions.

Let us conclude. We have established that there are unnameable propositional functions but we have also seen that this does not amount to saying that no significant facts about such functions are expressible. They can well partake in extensional talk. However, if we restrict ourselves to $\mathcal{L}_{\mathrm{RTT+Val}}$ sentences not exceeding a certain specific order n—for instance, along the way suggested at the end of section 2.6—we will realize that while we are able to express many facts regarding concepts of orders $\leq n$, there are facts that are expressible only if we permit ourselves to release the restriction and admit concepts of order $n + 1$. Indeed, as the statement of the very existence of such facts requires concepts of order $> n$, we cannot even recognize that there is something we have failed to express. To realize that, we have to make a step upwards and add variables of higher r-types. And this procedure can be iterated.

What causes this peculiar never-ending spiral of failure to express everything that, as if, should have been expressible? I propose that we should view this issue in terms of self-reproductive concepts identified in section

2.3. Both Tr and Het are such concepts. If this is so, the ramified theory of types (with the axiom of reducibility) appears to have achieved exactly what it was intended to. It permits self-reproductive concepts to be employed substantively in language, while blocking the paradoxes by not letting the classes determined by the corresponding predicative functions t and h form completed unities. This is surely a great merit deserving to be emphasized: Russell's ramified theory of types does not deal with the "rogue" concepts simply by shutting the door and pushing them outside the system. It lets them in, and permits to make a good use of them; they are just stripped off the "vicious" features that would bring in contradictions.

Chapter 3

Zermelo: Hierarchy of Sets

How can truth be defined within set theory? The primary goal of chapters 3 and 4 is to answer this question. In the present chapter, we will describe in some detail Zermelo's cumulative hierarchy of sets. This task is rather complex and requires introduction of a number of special concepts such as well-foundedness, being an inaccessible ordinal or quasi-categoricity. Close attention will be devoted to the criticism coming from Skolem since it threatens to break the whole Zermelian set-theoretic project down. Above all, though, we will strive to draw a coherent picture of a particular, and in some ways rather unorthodox, conception of set theory and the set-theoretic universe. In chapter 4, we will get to the business of finding, within set theory, an acceptable definition of set-theoretic truth. An attempt will be made to combine the hierarchical approach underlying Zermelo's conception of set theory with the findings resulting from the appreciation of some peculiar features of the concept of truth.

When Zermelo, more than 20 years after the publication of his famous paper containing the first axiomatization of set theory (Zermelo [1908]), returned to the foundational issues in the article 'Über Grenzzahlen und Mengenbereiche. Neue Untersuchungen über die Grundlagen der Mengenlehre' (Zermelo [1930]), his focal point was not the actual development of an axiom system itself but rather the study of the structures satisfying the axioms. The axiom system considered consists of the following standard first-order axioms of Zermelo-Fraenkel set theory:

- extensionality,
- pairing,
- union (sum),
- power set.

These axioms, in which only the variables ranging over sets occur, are supplemented by three axioms that contain also higher-order variables, namely:

- separation (often referred to by its German name 'Axiom der Ausson-

derung'),
– foundation (regularity),
– replacement.

The axiom of choice is not listed among the axioms because Zermelo takes it to be 'a general logical principle' (Zermelo [1930], p. 1220), i.e., a principle of wider generality, underlying the very logic of the language in which set theory is formulated. A notable thing is Zermelo's omission of the axiom of infinity, which is normally included among the standard axioms of set theory. The reason is that the axiom of infinity, without which the existence of infinite sets cannot be proved, is seen by Zermelo as a special strengthening of the underlying, ontologically lighter, set theory. Therefore, its adoption makes the resulting set theory less general.[1]

The theory Zermelo presents is second-order. This has the convenient property that the last three of the axioms stated above can be formulated as single axioms, not as schemata, as is the case in first-order set theory. To simplify the matter as much as possible, I will include in the theory considered in the present section the axiom of infinity, which Zermelo leaves out. The omission of the axiom has certain consequences that will always be explicitly mentioned—viz. p. 62 and the discussion at the end of section 3.6—so our simplification should not lead to any confusion or erroneous conclusions. I will refer to the resulting axiom system as to second-order Zermelo-Fraenkel set theory with the axiom of choice, ZFC_2. For the standard first-order theory I will use the abbreviation 'ZFC'.[2]

3.1 Well-foundedness

Among the axioms there is one that deserves a special attention, namely the axiom of foundation. With a certain degree of simplification, it can be said that this axiom was a novelty.[3] It aims at excluding circular or non-grounded sets by means of ensuring that the membership relation has

[1] Cf. Zermelo [1930], p. 1219. We will touch upon the issue of Zermelo's striving for utmost generality in section 3.6.

[2] To be more precise, some of the axioms listed above can be shown to be redundant as they are provable from the other axioms of ZFC or ZFC_2. The dispensable axioms are: the axiom (schema) of separation, which can be derived from the axiom (schema) of replacement, and the axiom of pairing, obtainable from the axioms of extension, replacement and power set. Still, they are, for the sake of comfort, usually listed among the axioms ZFC and ZFC_2. Besides, we will see shortly that there is also a first-order version of the axiom of foundation. This latter version is sufficient and may be kept without change in ZFC_2 since the second-order version of the axiom of foundation is provable from the first-order version together with the second-order axiom of replacement.

[3] The concept of well-foundedness can be traced back to Mirimanoff [1917], the discussion of an axiom ruling out the infinite descending chains can be found in Skolem [1923], von Neumann [1925] and von Neumann [1929], but Zermelo was the first to actually include the requirement of well-foundedness in the axiom system itself.

a special property called 'well-foundedness'. A (binary) relation R on a set a is WELL-FOUNDED if there is no infinite sequence $x_1, x_2, \ldots, x_n, \ldots$ of members of a such that $R(x_1, x_0), R(x_2, x_1), \ldots, R(x_{n+1}, x_n), \ldots$ all hold. This is often expressed by saying that R is well-founded if there are no infinitely descending R-sequences or R-chains. In particular, the membership relation is well-founded if there is no infinite $\in$-chain such that $\ldots \in x_{n+1} \in x_n \in \ldots \in x_1 \in x_0$. In a derivative way, a set is said to be well-founded if the membership relation on this set is well-founded. Zermelo's axiom of foundation is then nothing else than the requirement that every (descending) $\in$-chain, in which each term is a member of the preceding term, be of a finite length.[4]

However, in standard ZFC it is more common to define well-foundedness differently; the axiom then takes a different shape, too. The alternative definition goes as follows: the relation R on a set a is well-founded if every non-empty subset x of a has an R-minimal member, i.e., if there is a $z \in x$ such that $\neg R(y, z)$ for every $y \in x$. In particular, the membership relation on a is well-founded if every non-empty subset x of a has a member which itself does not have any members (in x). Reflecting this definition of well-foundedness, the alternative axiom of foundation assumes the following form:

$$\forall x\big(x \neq \varnothing \rightarrow \exists y[y \in x \wedge \forall z(z \in y \rightarrow z \notin x)]\big), \qquad \text{(A-Fou)}$$

i.e., every non-empty set x has some member y such that the intersection of x and y is empty.[5] Zermelo claimed that the two versions of well-foundedness we have just described were equivalent. Yet, this is not so generally. It has been proved by Mendelson [1958] that in order to get the latter version (A-Fou) from the former, we need the axiom of choice. Of course, we have said that Zermelo assumed the axiom of choice so, in his system, this condition is fulfilled; obviously, the same goes for ZFC_2.

To see that the version (A-Fou) of the axiom of foundation rules out the infinite (descending) $\in$-sequences is easy. Assume, as a counter-example, that there is a sequence $x_0, x_1, \ldots, x_n, x_{n+1}, \ldots$ such that $x_{n+1} \in x_n$ for each n, i.e., that there is an infinite (descending) sequence $\ldots \in x_{n+1} \in x_n \in \ldots \in x_1 \in x_0$. Then there is a set a containing all the terms of this sequence. Now apply the axiom (A-Fou) to a. The set a has to have a member, say x_k, which is disjoint from a, i.e., $x_k \cap a = \varnothing$. However, x_k and a have always a

[4] Cf. Zermelo [1930], p. 1220.

[5] This is a so-called 'local' version of the axiom of foundation. The axiom is sometimes expressed 'globally' using classes, $A \neq \varnothing \rightarrow \exists y(y \in A \wedge \forall z(z \in y \rightarrow z \notin A))$, or as a schema, $\exists y(\varphi(y)) \rightarrow \exists y(\varphi(y) \wedge \forall z(z \in y \rightarrow \neg\varphi(z)))$, where z is not free in φ. If $\varphi(y)$ is taken to be $y \in x$, the local version is equivalent to an instance of the schema. Note that, in contrast to the other axiom schemata, if we assume the so-called 'minimal member principle' for well-founded relations, which can be proved without the use of the axiom of foundation, the schema of foundation can be shown to be equivalent to its local instance. See Lévy [1979], p. 72, for details.

member in common, namely x_{k+1}. Therefore, the existence of the sequence assumed above is in contradiction with the axiom of foundation (A-Fou). How does the axiom preclude the existence of cyclical sets, i.e., sets that are members of themselves? If a set contains only itself, $a = \{a\}$, this is obvious as, in such a case, $a \cap \{a\} = a$. Yet if the set contains more elements than just itself, the axiom of foundation eliminates it only in conjunction with the axiom of pairing, according to which, for any two sets, there is always a set containing just these two sets and nothing else. Since a singleton is considered to be a special, degenerate case of a pair, we can conclude that for any set a there is always a set $\{a\}$. Now if a contains itself, the existence of the set $\{a\}$, guaranteed by the axiom of pairing, is in contradiction with the axiom of foundation (A-Fou). An analogical procedure can be applied to exclude sets that contain themselves not directly but are nested deeper inside their members.

Zermelo did not think that the axiom of foundation had the same status as the remaining axioms. It is true that, intuitively, it seems rather plausible but one thing is our inability to conceive of the possibility of the existence of non-well-founded sets and another thing is to say that they really are impossible.[6] Zermelo himself did not attempt to present any argument for the truth of the axiom; in fact, he thought that it represented a *restriction* imposed on the set-theoretic universe, so it can be claimed with good reason that he did not believe that the axiom was true.[7] The axiom of foundation is thus to be more properly viewed as a decision to circumscribe the universe of sets rather than a statement of a descriptive truth about sets. At the same time, despite its somewhat ambiguous status, the inclusion of the axiom of foundation among the axioms of set theory represents no imminent danger. It was proved already by von Neumann [1929] that adjoining the axiom of foundation to the rest of the axioms does not make the resulting theory inconsistent provided the original axioms were not inconsistent.

In justifying the axiom, Zermelo satisfies himself with a practical reason, noting that it 'has always been satisfied in all practical applications of set theory' so, in spite of its being of a restrictive nature, it does not restrict the theory in any essential way after all—not at least 'for the time being' (Zermelo [1930], p. 1220). This is, indeed, correct. It can be shown that arithmetic, analysis and virtually any branch of mathematics can be fully developed with the axiom of foundation. At the same time, however, work in all these disciplines can also be carried out without the axiom. The axiom

[6] Cf. Parsons [1977], p. 296. For instance, Suppes [1960], p. 53, satisfies himself with challenging the reader who does not find the idea of there being two distinct sets a and b such that $a \in b \wedge b \in a$ counterintuitive to give an example of these sets that would satisfy this condition. The task of constructing non-well-founded sets and, in general, of developing non-well-founded set theory by employing a so-called 'anti-foundation' axiom instead of the axiom of foundation has been most influentially taken up by Aczel [1988].

[7] Cf. Lavine [1994], p. 135.

of foundation is precisely the one axiom that is *not* necessary for doing mathematics in set theory, i.e., all the substantial mathematical results can be obtained without it. In this sense, it can be said that the notion of well-foundedness is characteristic of set theory proper as an autonomous field of study. As Kanamori [1996], p. 28, puts it, 'current set theory is at base the study of well-foundedness, the Cantorian well-ordering doctrines adapted to the Zermelian generative conception of sets'. But what is the point of indulging in a study of a notion that does not seem to have any significant practical implications for the very subject matter we are trying to understand, i.e., mathematics? The answer to this question is that well-foundedness becomes significant from a metatheoretical perspective, in an analysis of the structures satisfying the axioms.

However, it is necessary to add an important qualification. A metatheoretical study of structures is also a field where the question of the order of the theory comes to have decisive consequences. That is, it turns out that well-foundedness cannot be captured in first-order set theory in the sense that the membership relation would be well-founded in all the models satisfying the axioms. Despite the presence of the axiom of foundation, there will always be models of ZFC containing non-well-founded sets. How can this be? It is a corollary of the compactness theorem for first-order logic that a theory which has a model involving sequences of any finite length also has a model containing a sequence of an infinite length.[8] It was already pointed out by Skolem [1923], pp. 298–299, that the finitude of an R-chain is a higher-order property which cannot be successfully enforced in first-order ZFC since ZFC lacks the resources to exclude *all* infinite (descending) membership chains. In fact, it is able to exclude only those infinite chains that are definable in ZFC.

As we have seen, the axiom system presented in Zermelo [1930] is second-order. Although Zermelo does not present his axioms formally so it cannot be read off the axioms themselves, he states that the functions involved in the axioms of separation and replacement are to be read as 'quite *arbitrary*', without being in any way restricted (Zermelo [1930], p. 1220). So there is textual evidence for the claim that what he had in mind was a full second-order formulation of the axioms (i.e., with standard semantics).[9] Indeed, the

[8]For a discussion of the second-order character of well-foundedness see Shapiro [1991], pp. 108–109 and 113–114, and Kolman [2008], pp. 426–427.

[9]However, as Tait [1998*b*], p. 469n., points out, the question about what Zermelo intended is obscured by the fact that, immediately after asserting the arbitrariness of the aforementioned functions, he refers to his article on the concept of 'definiteness' (Zermelo [1929]) in which he identifies the definite functions with those definable in second-order set theory. If his arbitrary functions were supposed to be taken from among the definite functions, Zermelo's subsequent results would become invalid since they are only valid if we assume arbitrary functions without any restriction. Ebbinghaus [2007], pp. 183–186, depicts a gradual shift in Zermelo's conception of definiteness leading to the identification of this notion with categorical definability, and eventually to its absorption into the notion

textual evidence notwithstanding, the main reason for the full second-order reading of the axioms is the fact that, following the impossibility of enforcing well-foundedness in a set theory formulated in the first-order language, Zermelo's most significant results concerning the question of categoricity of the proposed axiom system, about which we will speak shortly, are not valid for first-order set theory. Hence, in what follows, we will be working with (second-order) ZFC_2 with standard semantics.

3.2 The Hierarchy and Inaccessible Ordinals

As we have seen, the axiom of foundation demands that the $\in$-relation be well-founded; by imposing a restriction on the notion of set it globally characterizes the set-theoretic universe. A straightforward consequence of the axiom is that all sets can be stratified into a hierarchy based on the length of the appropriate $\in$-chain of their elements. As all the descending $\in$-chains eventually lead to the empty set or to some primitive base-level objects without members that are not sets, so-called 'urelements', they can be compared, figuratively speaking, with respect to the number of "steps" needed to get all the way down to the base level. Accordingly, sets may be viewed as forming a hierarchy which can be represented as a (well-ordered) sequence of disjoint layers (Zermelo uses the word 'Schichten'). The hierarchy is cumulative in the sense that the sets in any given layer may take their members from any of the layers preceding the layer into which they themselves belong but not from this layer itself or any succeeding layers.

We shall now specify the cumulative hierarchy more precisely. To do this, we need to introduce the concept of von Neumann ordinal and the rank function. Zermelo called von Neumann ordinals 'basic sequences' ('Grundfolgen'), without giving any credit to von Neumann [1923] where they were introduced.[10] We will stick to the standard usage; but we will often abbreviate and write 'ordinals', meaning 'von Neumann ordinals'. A VON NEUMANN ORDINAL α is defined as a well-ordered set of the smaller von Neumann ordinals. Von Neumann ordinals form a sequence constructed as follows: $\varnothing = 0$; the successor of any ordinal α is $\alpha \cup \{\alpha\}$; if b is a set of ordinals, the union $\bigcup_{a \in b} a$ is an ordinal. Whenever a von Neumann ordinal α is smaller than a von Neumann ordinal β, then $\alpha \in \beta$ and $\alpha \subseteq \beta$, i.e.,

of set. To illustrate Zermelo's stance on the issue of arbitrariness, Ebbinghaus (op. cit.), p. 191, adds a quotation from a manuscript written in 1931 in which Zermelo argues against any attempt to restrict the scope of universality of the quantification over subsets of a given set. Hence, it seems reasonable, at least for our purposes, to view Zermelo's footnote as misleading and to read the requirement of arbitrariness as unrestricted. For a further discussion of the higher-order character of Zermelo's theory, see, for example, Hallett [1984], pp. 266–269, or Lavine [1994], p. 136.

[10] It is believed that Zermelo defined the sequence of von Neumann ordinals independently already around 1915 but did not publish his definition (cf. Ferreirós [2007], p. 376).

von Neumann ordinals are transitive. (A set a is TRANSITIVE if, for every $x \in a$, $x \subseteq a$ or, equivalently, if $y \in x$ and $x \in a$ imply that $y \in a$.) It can be shown that every well-ordered set is order-isomorphic to exactly one von Neumann ordinal, i.e., there is a bijection f from the well-ordered set a to a von Neumann ordinal β such that, for every $x, y \in a$, $x \leq y$ if and only if $f(x) \leq f(y)$. In this way, rather than being *identified* with equivalence classes of order-isomorphic sets—and that was, as we saw on p. 19, how ordinal numbers were defined by Cantor, whose definition was also taken up by Russell—von Neumann ordinals canonically *represent* such equivalence classes.

It remains to introduce the rank function. The RANK of a set is the von Neumann ordinal determining the level of the hierarchy at which the set first appears. It can be defined within ZFC as: $\rho(x) = \bigcup\{\rho(y) + 1 \mid y \in x\}$.[11] It can be shown that the rank is defined for all sets; that the rank of an ordinal is the ordinal itself; and that if $x \in y$, then $\rho(x) < \rho(y)$.

Now we are ready to describe the hierarchy itself. Take U to be a (possibly empty) set of urelements;[12] then the cumulative segments ('Abschnitte') of the hierarchy V_α, where α designates the appropriate rank, can be defined by transfinite induction in the following way:

$$\begin{aligned} V_0 &= U \cup \varnothing; \\ V_{\alpha+1} &= V_\alpha \cup \wp(V_\alpha); \\ V_\delta &= \bigcup_{\alpha<\delta} V_\alpha \text{ for limit ordinals } \delta. \end{aligned}$$

Ultimately, the entire universe V of sets is represented by the union of all segments:

$$V = \bigcup_\alpha V_\alpha.$$

It should be noted that while this specification of the hierarchy presented by Zermelo assumes urelements, in contemporary set theory it is customary to keep urelements out of the picture and deal exclusively with pure sets.[13]

[11] Cf. Drake [1974], pp. 35–36.

[12] Note, however, that no reason has been provided for the contention that the base, i.e., the collection of urelements, forms a set. As this issue is rather remote from the topic of this chapter, I will avoid discussing it; I would just like to mention the so-called 'urelement set axiom' introduced by McGee [1997], used to derive McGee's 'categoricity theorem'. For a further treatment, see Uzquiano [2002].

[13] Again, as with Zermelo's leaving out the axiom of infinity, what is at issue here is generality. A theory capable of coping with urelements as well as with pure sets is clearly more general than a theory capable of handling pure sets only. Indeed, as Kreisel [1967], p. 147n., points out, 'the classical structures of mathematics occur already, up to isomorphism, in the cumulative hierarchy *without* individuals [that is, urelements].' Therefore, Zermelo's interest in keeping set theory maximally general must have its roots elsewhere than in purely mathematical considerations. We will return to this issue in section 3.6.

The specification of segments V_α remains valid if we disallow urelements but it can be simplified: the first line becomes just $V_0 = \varnothing$ and the second line becomes $V_{\alpha+1} = \wp(V_\alpha)$.[14]

Each ordinal α indexes a level of the hierarchy. Some of these levels V_α have a rather special status, namely those that are indexed by what Zermelo called 'boundary numbers', and what is today referred to as '(strongly) inaccessible numbers'.[15] In brief, the segments indexed by these numbers are particular because they provide models for Zermelo-Fraenkel set theory; however, before going into the enquiry into models of ZFC_2, we need to understand what these numbers are. Unfortunately, the path leading to the definition of a strongly inaccessible ordinal is rather complex, and it is necessary to start off with several preparatory definitions. In particular, we need to understand the notions of being a (strong) limit ordinal and being regular. Furthermore, the latter presupposes the concept of cofinality.

Let us begin with the property of being a limit. An ordinal number $\alpha > 0$ is a LIMIT ORDINAL if there is no ordinal number β such that $\beta + 1 = \alpha$, i.e., if it cannot be reached by the successor operation. An ordinal number $\alpha > 0$ is a STRONG LIMIT ORDINAL if there is no ordinal number β such that $2^\beta = \alpha$, i.e., if it cannot be reached by the power-set operation. Note that every strong limit ordinal is, at the same time, a limit ordinal. The other way round, however, the matter is less straightforward. In order to claim that every limit ordinal is also a strong limit ordinal, we would have to show that there are no limit ordinals between ω and 2^ω, in particular, and between ω_α and 2^{ω_α} for every ordinal number α, in general. And this amounts to nothing else than to Cantor's continuum hypothesis, i.e., to the statement that $\aleph_1 = 2^{\aleph_0}$, and to the generalized continuum hypothesis (first conjectured by Hausdorff [1908]), which is the statement that $\aleph_{\alpha+1} = 2^{\aleph_\alpha}$, respectively. Thus, if the generalized continuum hypothesis is assumed to hold, strong limit ordinals are precisely limit ordinals.

The next auxiliary notion we need to introduce is that of cofinality. A function $f : \beta \mapsto \alpha$ is COFINAL (in α) if the image $f[\beta]$ is unbounded in α, i.e., if $\forall \xi \in \alpha \; \exists \eta \in \beta (\xi \leq f(\eta))$. The COFINALITY $\mathrm{cf}(\alpha)$ of an ordinal number α is then defined as the smallest ordinal number β such that there is a cofinal function $f : \beta \mapsto \alpha$, i.e., $\mathrm{cf}(\alpha) = \min\{\beta \mid \text{there is a cofinal function } f : \beta \mapsto \alpha\}$. The notion of cofinality deserves several brief remarks.[16] First,

[14] The simplification of the second line is based on the fact that, disallowing urelements, it can be shown that, for any ordinal α, V_α is a transitive set (cf. Balcar and Štěpánek [2000], p. 193).

[15] The theory of inaccessible numbers, which plays a key role here, was developed by Hausdorff [1908]. Hausdorff used the term 'exorbitant' instead of 'inaccessible', and applied it to what would now be called 'weakly inaccessible' numbers. Hallett [1996], p. 1210n., points out that the term 'inaccessible' was probably first introduced by Sierpinski and Tarski [1930] who used it for what would be now called 'strongly inaccessible' numbers.

[16] A more complete treatment of the concept of cofinality can be found, for example, in

the cofinality of any ordinal number α is less than or equal to α since the identity function is cofinal. Secondly, $\mathrm{cf}(0) = 0$. The cofinality of every successor ordinal is 1 since there is a function $f : 1 \mapsto \alpha + 1$ such that $f(0) = \alpha$ which is cofinal in $\alpha + 1$. If $\mathrm{cf}(\alpha) = 1$, α has a greatest element since there is always a cofinal function (in α) from 1 to this element. The cofinality of a limit ordinal is $\geq \omega$; in particular, if $\alpha < \omega_1$, then $\mathrm{cf}(\alpha) = \omega$. An ordinal number α is called SINGULAR if $\mathrm{cf}(\alpha) < \alpha$; if, on the other hand, $\mathrm{cf}(\alpha) = \alpha$, the ordinal number α is said to be REGULAR. The only finite numbers that are regular are 0 and 1; all the other natural numbers are singular. ω is regular. If we follow the usual path and accept the axiom of choice, any infinite successor ordinal $\omega_{\alpha+1}$ is regular. Every infinite ordinal that is regular is also a so-called INITIAL ordinal, i.e., an ordinal number such that every smaller ordinal has a smaller cardinality.[17] But not every initial ordinal is regular since there are also singular initial ordinals: a limit of regular ordinals is an initial ordinal although it is typically not regular, for instance, ω_ω whose cofinality is merely ω. A characteristic property of a singular infinite initial ordinal α (which is often used as an alternative definition of singularity) is also that α can be represented as the union of $< \alpha$ sets each of which is of cardinality $< \alpha$. Assuming the axiom of choice, this is equivalent to the condition that α can be represented as the sum of $< \alpha$ cardinals each of which is $< \alpha$.[18] If no such representation is possible, the ordinal α is regular.

To sum it up, ω and the successor ordinals are regular but limit ordinals ω_α with α a limit ordinal are typically singular. Yet what does the 'typically' mean? Are there, besides ω and the regular successor ordinals, also regular *limit* ordinals? An ordinal that satisfies the properties just mentioned, namely:

- is $> \omega$ (alternatively, is uncountable),[19]
- is regular and
- is a limit ordinal,

is called a WEAKLY INACCESSIBLE ordinal. If we strengthen the third condition imposed on the ordinals from being a limit ordinal to being a strong

Lévy [1979], pp. 133–141, and Ciesielski [1997], pp. 74–76.

[17] Provided that every set can be well-ordered, which follows from the axiom of choice, every cardinal has an initial ordinal. It is, therefore, customary to identify cardinal numbers with their corresponding initial ordinals. As both strongly and weakly inaccessible ordinals defined below are always initial ordinals, they can be identified with their corresponding cardinals; in fact, they are usually presented as inaccessible cardinals. However, with the intention of making the matter as simple as possible and avoiding introducing cardinal numbers, we will stick to speaking of inaccessible ordinals.

[18] For proofs of the equivalence of these two conditions with the defining condition of regularity, see Lévy [1979], pp. 134–135.

[19] The additional condition of uncountability is there to exclude ω from the inaccessible ordinal numbers. The issue of ω being the first inaccessible ordinal in Zermelo's system is discussed below.

limit ordinal, we get a (STRONGLY) INACCESSIBLE ordinal (hereafter only an 'inaccessible ordinal'). Equivalently, a weakly or strongly inaccessible ordinal is an ordinal that cannot be reached 'from below' by the general cardinal addition as well as (if it is weakly inaccessible) by the successor operation, i.e., by passing from ω_α to $\omega_{\alpha+1}$, or (if it is strongly inaccessible) by the exponential operation, i.e., by passing from ω_α to 2^{ω_α}. Indeed, as we have already pointed out, if we assume the generalized continuum hypothesis, these two notions coincide.

Our definition of the notion of inaccessible ordinal is now complete. It is clear that if such numbers exist, they have to be enormous, much larger than the ordinals that arise in ordinary mathematical practice (cf. Jech [1991], p. 42). Yet the question that needs to be asked is: Do such numbers exist? It is certainly conceivable that inaccessible ordinals exist, but do they? We will deal with this question and its consequences in the next section; for the time being, let us just say that the existence of uncountable inaccessible ordinals cannot be proved in ZFC or ZFC_2.

Note Zermelo's conception of the cumulative hierarchy of sets is sometimes put in connection with the so-called 'iterative conception of set'. The iterative conception was first clearly described in print in Gödel [1947], pp. 473–477, and later it was developed in Shoenfield [1967], pp. 238–240, Boolos [1971], Scott [1974] and Wang [1974], pp. 181–193. The primary goal of the iterative conception is to provide an independent and intuitively acceptable justification for the axioms of set theory via supplying an intended, standard model. On its basis, systems of set theory such as ZFC can be defended against the accusation that they are ad hoc—built chiefly as a response to the threat of the set-theoretic paradoxes—and lack any serious intuitive or philosophical appeal.[20] The central metaphor of the iterative conception is that, at any moment, we can produce new sets by gathering together objects that we already have at our disposal. Then, by iterating the operations of creating new and new layers of sets on top of those already present, we usually arrive at the familiar cumulative hierarchy of Zermelo's. However, the cumulative hierarchy seems to be all that Zermelo [1930] and the iterative conception have in common. Zermelo's article is above all a study in models satisfying the axioms of ZFC_2, and not an attempt to philosophically justify the axioms by building on our ontological or metaphysical

[20] Boolos claims that the iterative conception is 'natural' (cf. Boolos [1989], p. 89, or Boolos [1998*b*], p. 127) in the sense that 'without prior knowledge or experience of sets, we can or do readily acquire the conception, easily understand it when it is explained to us, and find it plausible or at least conceivably true'. However, the assertion that the conception is natural has recently been subjected to criticism. A challenge has appeared in the development of non-well-founded set theory (cf. Aczel [1988]) and in the attempts to show that the iterative conception can be relatively easily modified to justify also non-well-founded universes of sets (see, e.g., Sharlow [2001]).

intuitions. As Taylor [1993], p. 553, argues, if understood in this way, Zermelo [1930] does not develop the iterative conception, and should not be seen as its forerunner.

3.3 The Sequence of Models

We have said that each ordinal α indexes a segment within the cumulative hierarchy described in section 3.2, and that among segments there are some which are particularly significant, namely those indexed by inaccessible ordinals. These segments provide models of Zermelo-Fraenkel set theory. We are now ready to address this issue in a slightly greater detail.

In general, a MODEL, or a STRUCTURE, is a complex object consisting of a domain, which is usually taken to be a set of objects, and an interpretation function that assigns appropriate objects from the domain of the model to the primitive non-logical vocabulary of the language of the theory, i.e., to the individual constants, relation symbols and function symbols. In the particular case of set theory, a model will contain a domain together with the interpretation of the single primitive relation symbol, '$\in$'. If $\in_a$ is the set of all ordered pairs $\langle x, y\rangle$ such that both x and y are members of a and $x \in y$, and V_α is a segment within the cumulative hierarchy defined in section 3.2 with α being an inaccessible ordinal, then the ordered pair $\langle V_\alpha, \in_{V_\alpha}\rangle$, which I will designate simply by '$\mathcal{M}_\alpha$', can be shown to be a model of ZFC or ZFC_2.[21] Let us call models of this form NORMAL MODELS. (When Zermelo [1930], pp. 1220, 1224, speaks of objects satisfying his axiom system with respect to the $\in$-relation, he speaks of 'domains'—which he calls 'normal domains'—, not of 'models'. In what follows, I will stick to 'normal models' to avoid confusion of models and domains.)

Of course, the segments of the cumulative hierarchy indexed by inaccessible ordinals are not domains of models of ZFC_2 by accident. It can be shown that a segment's being indexed by an inaccessible ordinal is a condition that is both necessary and sufficient for its being the domain of a normal model of ZFC_2. It is *sufficient* if it can be shown that the domain of a normal model contains (at least) every set whose existence can be proved in ZFC_2. It is straightforward to see that this requirement is satisfied by inaccessible segments. If α is a limit ordinal, then V_α already provides for all the sets provable by the axioms of extensionality, separation, pairing, union, power set and foundation. If $\alpha > \omega$, V_α provides for the axiom of infinity.[22] It remains to show that V_α also contains all the sets obtainable by the axiom

[21] We have already mentioned the fact that if the axiom of infinity is removed from ZFC or ZFC_2, the existence of ω becomes unprovable. If, moreover, we drop, as Zermelo does, the condition of being greater than ω from the definition of the inaccessible ordinal, ω will come out as the first inaccessible ordinal. Then the normal model $\mathcal{M}_\omega$ will be a model of such a 'finitistic' version of set theory.

[22] For the detailed proofs, see Jech [2003], pp. 165–166.

of replacement, which says, in the extended notation, that

if f is any function, then for every set z the image $f[z]$ is a set. (A-Rep)

To show that the axiom of replacement does not lead us outside the segments indexed by inaccessible ordinals, assume that there is a set $z \in V_\alpha$ and a function $f : z \mapsto V_\alpha$, with α an inaccessible ordinal. The following facts may be stated: (a) for any $z \subseteq V_\alpha$, $z \in V_\alpha$ if and only if $|z| < \alpha$; and (b) the cardinality of the image $f[z]$ is less than or equal to that of z, i.e., $|f[z]| \leq |z|$. Thus we get $|f[z]| \leq |z| < \alpha$, from which it follows by (a) that $f[z] \in V_\alpha$.[23] Therefore, it can be concluded that a segment indexed by an inaccessible ordinal contains all the sets obtainable by the use of the operations of set formation sanctioned by the axioms of ZFC_2; hence the condition of being indexed by an inaccessible ordinal is, indeed, sufficient.

The same condition is *necessary* if it can be ruled out that some other domains, smaller than the segments indexed by inaccessible ordinals, can serve equally well as models of ZFC_2. This is to say that we need to show that ZFC_2 is satisfied by only those segments indexed by ordinals that are $> \omega$, regular and strong limit. We may skip the property of being $> \omega$ as it has already been discussed, and go straight to regularity. Assume that V_α is a domain of a model of ZFC_2 and that α is singular. Then there will be an ordinal $\beta < \alpha$ and a cofinal function $f : \beta \mapsto \alpha$ (with the image $f[\beta]$ unbounded in α) such that $f \subseteq V_\alpha$. Hence the supremum of $f[\beta] = \alpha$. However, by the axiom of replacement we get that $f[\beta]$ is a set, i.e., $f[\beta] \in V_\alpha$, from which it follows that the supremum of $f[\beta]$ also has to belong to V_α, i.e., $\alpha \in V_\alpha$. And that is a contradiction. It remains to show that α has to be strong limit. Again, assume that it is not. Then there will be a $\beta < \alpha$ such that $|\wp(\beta)| = 2^\beta \geq \alpha$, and, obviously, by the power set axiom, the power set $\wp(\beta)$ is a set, i.e., $\wp(\beta) \in V_\alpha$. Yet, there will be a surjective function $f : \wp(\beta) \mapsto \alpha$ such that the image of $\wp(\beta)$ under f is equal to α, i.e., $f[\wp(\beta)] = \alpha$. But then, by the application of the axiom of replacement, also the image $f[\wp(\beta)] \in V_\alpha$, from which it follows that $\alpha \in V_\alpha$. And that is a contradiction.[24] It is important to observe that the necessity of the condition cannot be established without crucial application of the full second-order axiom of replacement (A-Rep). This entails that there are models of (first-order) ZFC whose rank is smaller than the smallest inaccessible ordinal.[25]

We have established that any segment V_α with α an inaccessible ordinal provides a model of ZFC_2. But certainly, there will be a multitude of

[23] This proof is taken from Kanamori [2003], p. 18.

[24] This proof is also taken from Kanamori [2003], p. 19. An informal argument in support of the necessity condition is already present in Zermelo [1930], pp. 1221–1224. The formal proof is attributed to Shepherdson [1952].

[25] This result was proved in Montague and Vaught [1959]. As they point out on p. 220, this is equivalent to saying that the axioms of ZFC do not insure the existence of all "accessible" ordinals.

different, non-isomorphic models. What is the relationship between them? What differences may one expect? One direction in which they may differ is obviously the cardinality of their base, i.e., the set of urelements. This has been called the "width" of the model. The other direction in which models may differ is their rank, that is, their "height". It is a chief result obtained by Zermelo that, in general, there can be no other point of difference between models of ZFC_2. Any normal model of ZFC_2 is uniquely specified (up to isomorphism) by a pair of characteristic cardinal numbers, namely by its height and by its width. Moreover, all normal models of equal width but different height have another particular property. As the hierarchy of sets is cumulative, i.e., each segment V_α contains all the segments with ranks $< \alpha$, models $\mathcal{M}_\alpha$ and $\mathcal{M}_\beta$ differing only by their ranks α and β differ (up to isomorphism) only in the layers situated, within the hierarchy, between V_α and V_β. A similar situation obtains with the cardinality of the base. Zermelo saw this result as a particular form of categoricity, sometimes called 'almost-categoricity' or 'quasi-categoricity': let us say that a theory is QUASI-CATEGORICAL if any domain (of a model satisfying it) with a lower height is isomorphic to a (proper or improper) subdomain of the domain with a higher height, and the domain with a lower width is isomorphic to a (proper or improper) subdomain of the domain with a higher width.

So, we have seen that the property of being indexed by an inaccessible ordinal α is both sufficient and necessary for V_α to be the domain of a normal model of ZFC_2, and that the models of ZFC_2 have a peculiar quality of being quasi-categorical. Yet, there still remains an important question to be touched upon, namely that of the existence of inaccessible ordinals. We claimed in the previous section that the existence of these ordinals is unprovable in ZFC_2. Why is this so? Once it has been established that the segments of the cumulative hierarchy indexed by inaccessible ordinals are domains of normal models of ZFC_2, the unprovability of the existence of these ordinals becomes immediate. For had ZFC_2 been able to prove the statement that there is an inaccessible ordinal α, it would have been able to prove the existence of its own model $\mathcal{M}_\alpha$, which amounts to nothing else than proving its own consistency. Of course, by Gödel's second incompleteness theorem, this is impossible unless the theory is inconsistent.[26] Therefore, the existence of inaccessible ordinals cannot be proved in ZFC_2.

However, this does not settle the question whether these ordinals do or

[26]The fact that provability of the existence of an unaccessible ordinal is inconsistent with ZFC_2 can also be established without recourse to Gödel's second incompleteness theorem, in a rather straightforward way. Assume that we are able to prove the existence of some inaccessible ordinals in ZFC_2. There will be the smallest such ordinal, call it 'θ'. We have shown that the segment V_θ will be a model of ZFC_2. Now, because θ is the smallest inaccessible ordinal there will be no inaccessible ordinal in V_θ. However, the existence of θ being provable, θ, an inaccessible ordinal, has to be in V_θ. And that is a contradiction. This argument goes back to Kuratowski [1925].

do not exist. We have merely found out that, if they exist, they cannot be shown to exist in ZFC_2 by means of a formal proof. Before going into the informal argument in support of their existence, let us point out that the situation we have here is, in a sense, remarkably similar to that of the question whether there are infinite sets in set theory lacking the axiom of infinity, which was, as we have seen, the system presented in Zermelo [1930]. Of course, the existence of ω is provable in ZFC. (Provided that we do not restrict the definition of inaccessible ordinals to those strictly greater than ω (or the uncountable ones), it can be proved that the statement that ω is the only inaccessible ordinal is consistent with ZFC. But this is merely a rewording of the fact that the existence of ω is provable in ZFC.) The reason for the provability of the existence of ω, however, lies nowhere else than in the presence of the axiom of infinity in ZFC. The axiom of infinity can be stated in the extended notation as the following statement:

$$\exists z(\varnothing \in z \wedge \forall x(x \in z \rightarrow x \cup \{x\} \in z)). \qquad \text{(A-Inf)}$$

In English, the axiom says that there is a set z such that it contains the empty set together with the union of each of its members with its singleton.[27] It is easy to see that this set contains all finite von Neumann ordinals, i.e., all natural numbers; therefore, it must be infinite. The axiom thus postulates the existence of an actually infinite set, and it does so without providing a recipe for its construction using the operations sanctioned by the other axioms. Take (A-Inf) out, and the existence of any infinite set will become unprovable. Moreover, if ZFC without the axiom of infinity is consistent, the negation of the axiom of infinity can be adjoined to the other axioms, and the resulting system will also be consistent. In this way we obtain a set theory equivalent to elementary number theory, i.e., arithmetic (cf. Jech [1991], pp. 38–39), in which all objects concerned are finite. Consequently, it can be said that the provability of the existence of ω in ZFC is achieved by our *decision* to add (A-Inf) to the remaining axioms, and thus to have ω among the provable sets. Indeed, this decision can be perfectly legitimate if there are good reasons justifying it (into which we will not go) but the fact is that it remains a decision. This stipulative character of the axiom of infinity is presumably the reason why Zermelo leaves it out from 'general' set theory.[28]

[27] This is the standard, most commonly used version of the axiom of infinity. For the comparison of the deductive strength of other, different versions of the axiom, including Zermelo's original one, see Uzquiano [1999], pp. 290–294.

[28] A word of caution is needed here. Kanamori [2004], pp. 524–525, points out that ZFC_2 without the axiom of infinity (A-Inf) and with only the local version of the axiom of foundation such as our (A-Fou) does not establish the theorem of transitive containment (which says that every set is a subset of a transitive set), without which the second-order form of the axiom of foundation cannot be derived. Such a second-order set theory thus has non-well-founded models. So if we want to preserve well-foundedness of all the models

Let us go back now to the question of the existence of inaccessible ordinals. What reasons, if any, can we put forward for their existence? If there cannot be a formal proof, is there a compelling informal argument? Before anything else, notice that if there are normal models whose domains contain infinite sets, i.e., if there is any "infinitistic" set theory at all, these models contain the sequence of all the ordinals whose existence can be proved in ZFC_2. To say that there is an inaccessible ordinal is then just another way of saying that there is an ordinal limit of this sequence of "accessible" ordinals, and that, in contrast to the "accessible" ordinals, this limit is not a set of ZFC_2. Still, why should we accept that there is a number greater than all the numbers that are provable in ZFC_2? What does Zermelo have to say? His argument for the existence of inaccessible ordinals (cf. Zermelo [1930], p. 1232) is based on the observation that the height of each natural model of ZFC_2 is uniquely fixed by the totality of inaccessible ordinals that it contains as members. This follows from the simple fact that the segment V_α indexed by an inaccessible ordinal α will contain all the ordinals, including the inaccessible ones, that are $< \alpha$. If V_α does not contain any inaccessible ordinals at all, α must be the smallest one; if it contains exactly one, α has to be the second smallest one, etc. As only the width of the domain is relevant for the question of the existence of inaccessible ordinals, let us deal for now only with models of equal width. Assume that there is an inaccessible ordinal α in the domain but no ordinal $> \alpha$, and let us add this assumption to ZFC_2 as a new axiom. Such an extended theory will be satisfied uniquely (up to isomorphism) by the model with the domain V_β with β the first inaccessible ordinal $> \alpha$. (If we were after fixing the domain indexed by the smallest inaccessible ordinal, we would have to append the axiom that there are no inaccessible ordinals at all.) In this way we are able to obtain from quasi-categorical ZFC_2—disregarding the urelements—a sequence of its categorical extensions. This entails, at the same time, that any domain whose rank is an inaccessible ordinal can be categorically specified.

Categoricity is a property of prime significance to Zermelo. He puts forward, as a 'general hypothesis', the following principle:

> [E]very categorically determined domain can also be interpreted [aufgefaßt] as a set in some way, i.e. can appear as an element of a (properly chosen) normal domain. (Zermelo [1930], p. 1232; Hallett's translation.)

To put it simply, the principle states that any collection of sets that can be categorically determined by means of a suitable extension of ZFC_2 is also a set. What is the justification for such a principle? It is rather simple:

satisfying the axiom system of ZFC_2 without the axiom of infinity, we have to have a full second-order version of the axiom of foundation. For a detailed analysis of the dependence of the property of well-foundedness of second order Zermelo set theory on various versions of the axiom of infinity, see Uzquiano [1999].

recall that, disregarding the urelements, ZFC_2 is quasi-categorical, i.e., it is satisfied exclusively by the sequence of models that differ just in their height. Moreover, as we have seen, models of greater height contain the models of smaller height as proper subsegments. Now, provided that any model can be uniquely fixed (up to isomorphism) by a categorical extension of ZFC_2, any model apart from the one with the smallest height will contain all the smaller inaccessible ordinals as elements. And as elements of the cumulative hierarchy, they will be ordinary sets. Without categoricity, there is no clear way of establishing whether a given inaccessible ordinal is or is not an element of the normal domain and, consequently, whether or not it can be treated as a set. This concludes the argument for the existence of inaccessible ordinals. A notable quality of the whole argument is, as Zermelo quickly adds (op. cit., pp. 1232–1233), that it provides us not merely with a single inaccessible ordinal but, in one sweeping move, with an unbounded sequence of ever greater inaccessible ordinals.[29] And naturally, given this unbounded sequence of inaccessible ordinals, there will be an unbounded sequence of larger and larger normal models of ZFC_2.

So we can now understand why, despite the fact that no inaccessible ordinal can be obtained from the base by means of the operations sanctioned by the axioms of ZFC_2, and despite the fact that no inaccessible ordinal is a set of ZFC_2, Zermelo makes a decision—an 'inspired move', as Kanamori [2004], p. 522, puts it—to take the number characterizing the height of the normal model to be again a standard (von Neumann) ordinal, subject to a proper set-theoretic treatment in an extension of ZFC_2.[30] By this move, the characteristic α of a normal model $\mathcal{M}_\alpha$ of ZFC_2 is kept outside the model but still within set theory.[31] Thus every single one of normal models can be fully specified set-theoretically, i.e., it can become an ordinary set capturable by the resources of set theory, provided that the existence of the requisite ordinal number is assumed. In contrast to this, however, the matter is entirely different as regards the whole universe of sets. If we want to capture a segment of the hierarchy as a set, we need to, so to speak, make a step upwards onto a higher layer. Obviously, this requirement makes capturing of the universe of all sets by means of any extension of ZFC_2 impossible. Or, as Feferman [1999], p. 101, puts it, since any closure condition on sets specifies a set, i.e., an object within the universe, the whole universe of sets cannot be captured by any set-theoretic closure. To sum it up, the cumulative hierarchy of sets contains an unbounded sequence

[29] In fact, one is not forced to stop with the sequence of inaccessible ordinals. Analogical methods may be applied in arguing for the existence of 'hyper-inaccessible' ordinals, defined as those inaccessible ordinals α that have α inaccessible ordinals below them. Segments indexed by hyper-inaccessible ordinals will be models of ZFC_2 extended by the axioms postulating ordinary inaccessible ordinals (cf. Tiles [1989], p. 181).

[30] Cf. also Ebbinghaus [2007], p. 190.

[31] Cf. Hallett [1996], pp. 1209–1210.

of higher and higher normal domains, each of which is capturable by a set-theoretic closure, except for the universe of sets, the hierarchy in its entirety, which remains, for any extension of ZFC_2, perfectly elusive.

3.4 The Challenge of "Skolemism"

Zermelo's efforts in the period in which his 'Grenzzahlen' article was published were focused most prominently on the struggle against the dangers of "Skolemism",[32] which he saw as threatening the very constitution of mathematics.[33] The pivotal role in "Skolemism" is played by the Löwenheim-Skolem theorem. To state this theorem in its most general contemporary form,[34] we need to be able to assess the cardinality of a language. By saying that a language has cardinality λ, we will mean that it contains λ symbols or, equivalently, that it has λ well-formed formulas (to be more precise, the equivalence holds for all languages containing infinitely many symbols, which requirement is satisfied, among other things, by all languages based on standard predicate logic, e.g., by the language of set theory). In particular, if the cardinality of a language is $> \aleph_0$, the language is said to be uncountable; otherwise it is countable. Now let T be a theory in a first-order language of cardinality λ. Then the DOWNWARD LÖWENHEIM-SKOLEM THEOREM says that if T has a model, it has a model with the domain of cardinality $\leq \lambda$. In particular, if T is a theory in a countable language, it has a countable model, i.e., a model with the domain of cardinality $\leq \aleph_0$. The UPWARD LÖWENHEIM-SKOLEM THEOREM says that if T has an infinite model, it also has a model of every cardinality $\kappa \geq \lambda$. Thus even if T is intended to deal exclusively with countably infinite sets, it still has models with domains of uncountable cardinalities.

It follows from the Löwenheim-Skolem theorem that formal theories attempting to characterize simple countable structures such as that of natural numbers are bound to characterize equally well also much more complex

[32] I use the double quotation marks to stress that what I am after is not the true nature of Skolem's ideas concerning set theory but rather just those aspects of the doctrines of Skolem's that, like a scarecrow, deter Zermelo from taking all the field of set theory as granted. For an elaboration of Skolem's views concerning the foundations of set theory, see Jané [2001].

[33] Cf. Ebbinghaus [2007], pp. 200–202.

[34] Cf. Enderton [2001], pp. 151–155. The history of the theorem is relatively complex. A version of the downward Löwenheim-Skolem theorem limited to a single formula was published in Löwenheim [1915]. Löwenheim's proof was reworked and extended to possibly infinite sets of formulas in Skolem [1920]. The upward part of the theorem has allegedly been proved by Tarski in 1928 but Tarski did not publish his result (cf. Moore [1982], pp. 257–258). Still, the upward part is sometimes called Löwenheim-Skolem-Tarski theorem. The theorem in its full generality was proved in Malcev [1936]. As Hodges [1997], p. 127, points out, it is somewhat paradoxical that Skolem's name is associated with the upward part of the theorem since 'Skolem didn't even believe it, because he didn't believe in the existence of uncountable sets'.

structures of huge uncountable cardinalities, while theories powerful enough to prove the existence of uncountable sets, e.g., of the continuum of real numbers, can also be satisfied by natural numbers. Not only that such theories are not categorical but their non-categoricity is such that it seems to undermine some of the most fundamental set-theoretic notions. Just recall that at the very heart of Cantor's set theory was its ability to distinguish different magnitudes of infinity. The Löwenheim-Skolem theorem then tells us that even though we can prove that there are sets of huge cardinalities that are uncountable within the "sandbox" of ZFC, what we have really achieved by such proofs is only that we have obtained new results characterizing, inter alia, the natural numbers. This perplexing phenomenon has been called 'Skolem's paradox'.

Skolem's own response to the challenge presented by this state of affairs was relativism according to which every 'thoroughgoing' axiomatization of set theory 'leads to a relativity of set-theoretical notions'. Specifically, 'higher infinities exist only in a relative sense' (Skolem [1923], p. 296).[35] Not surprisingly, such a position was utterly unacceptable for Zermelo, and his conception of the cumulative hierarchy may be viewed as a reply to this relativism of Skolem's.[36]

Now, at a first glance, it might look as if Skolem's relativism need not worry Zermelo at all since, as it is formulated in a full second-order language with standard semantics, ZFC_2 is immune to the Löwenheim-Skolem theorem in either direction. The significance of the requirement that Zermelo's theory should be understood to be second-order was already discussed in connection with the notion of well-foundedness. The immunity to the Löwenheim-Skolem theorem is just another feature characterizing a full second-order system; once a system of set theory becomes satisfiable only if all sets are well-founded, it is no longer affected by the Löwenheim-Skolem

[35] Just a word of caution. As Bernays [1957], p. 8, stresses, from this relativity it does not follow that one and the same set, e.g., $\wp(\omega)$, is uncountable in one theory and countable in another. Cantor's cardinality theorems are invariant with respect to different systems of axiomatic set theory; the set $\wp(\omega)$ is uncountable in *every* axiomatic set theory. The relativity lies rather in the fact that the totality of the subsets of ω that are definable in a given theory T_1 can be countable in a different, more comprehensive theory T_2. In T_2, however, this totality will no longer be the set $\wp(\omega)$.

[36] Here it should be noted, though, that Zermelo's understanding of the Löwenheim-Skolem theorem, its underlying causes and its consequences was problematic. It can be evidenced that Zermelo, who got into both professional and scientific isolation during the 1930s, did not fully catch on with the development in the field of logic and set theory, which shifted decisively towards first-order systems and a study of their different models. Ebbinghaus says: 'It is a tragic coincidence that Zermelo, although having contributed crucial model-theoretic features in the *Grenzzahlen* paper, completely missed this development' (Ebbinghaus [2007], p. 204). At the same time, one is most probably not justified in concluding that Zermelo simply missed the point of the Löwenheim-Skolem theorem. A more charitable reading of Zermelo's late attempts at a refutation of Skolem's relativism consists in realizing that in Zermelo's version of set theory 'no Skolem phenomenon could occur' (van Dalen and Ebbinghaus [2000], p. 158).

theorem. Thus the threat of a collapse of the set-theoretic universe and of its replacement by dissimilar models of all possible cardinalities has been staved off. Nevertheless, for the reasons that will appear shortly, Zermelo could not sensibly defend his view of set theory merely by pointing out that he was working out a second-order system of set theory. Skolem's critique cuts deeper.

When Cantor introduced uncountable sets, it was on the basis of his diagonal proof which established that no countable set of sets of integers can contain all sets of integers or, alternatively, that no countable set of subsets of a countably infinite set a can contain all the subsets of a. Yet, from this result itself it does not follow that there *is* an uncountable set of sets of integers or a set of subsets of a.[37] This requires an additional step which is provided for in ZFC by the power set axiom, according to which, for any set x, there exists a set $\wp(x)$ consisting of all the subsets of x. In the extended notation:

$$\forall x \exists y \forall z (z \in y \leftrightarrow z \subseteq x). \qquad \text{(A-Pow)}$$

However, in theories formulated in a first-order countable language, the power set axiom only establishes the existence of the set $\wp(x)$ of all the subsets of x that are first-order definable. Therefore, as the number of the first-order definable subsets is countable, such theories are satisfiable by countable models. This situation does change in theories formulated in higher-order languages with standard semantics, in which the quantifiers are assumed to range over all subsets of a domain, no matter whether second- or higher-order definable or not. It is only in such a higher-order framework that the power set axiom establishes the existence of "genuine" uncountable sets ('genuine' being used to indicate that theories containing theorems about such sets cannot be satisfied by countable models). Thus, in order to block the Löwenheim-Skolem theorem, one has to presuppose full higher-order logic, which amounts to assuming 'that if one countenances a domain d, one also countenances each subset of d, and each relation and function on d' (Shapiro [1991], p. 255). Unfortunately, this assumption appears to be somewhat problematic. We can certainly prove the existence of uncountable sets simply by requiring that the quantifiers of a given language range over uncountably many subsets of a given domain. Yet, this just appears to beg the question. Skolem may respond: When one is considering the possibility of introducing second-order variables ranging over arbitrary propositional functions, 'the question arises: what is the totality of all propositional functions?' (Skolem [1928], p. 516).[38] How can it be specified? The totality obviously cannot be fixed by a system of axiomatic set theory itself.

[37] Cf. Jané [2001], p. 130.

[38] Indeed, for Skolem, the only 'scientifically tenable' propositional functions are those that do not lie beyond the means available to a given first-order theory, namely those that are first-order definable. This is the position he assumed in Skolem [1923] in the

The whole argument against the existence of uncountable sets and the acceptability of the assumption of arbitrary subsets of a domain can be generalized and directed against the utility of the whole business of axiomatization of set theory. For it is axiomatic set theory that is supposed to determine or implicitly define the objects and their properties that satisfy the conditions it imposes, viz. its models. Yet, once it has been shown that the capability of axiomatic set theory of characterizing models is rather limited, the existence and properties of models simply need to be assumed beforehand. That is, it needs to be assumed that we are, in general, able to deal with models in a direct manner. But how are we supposed to do that? After all, 'models are usually taken to be set-theoretical objects' (Jané [2001], p. 142), and it would seem that if we want to deal with them, we need a proper set-theoretic framework, namely an axiomatic system. Hence, we have to conclude that in order to set up axiomatic set theory, we need to presuppose a general conception of model but, at the same time, in order to study models, we need to have a decent system of axiomatic set theory. Thus we seem to be trapped in a circle:

> If we adopt Zermelo's axiomatization, we must, strictly speaking, have a general notion of domain in order to be able to provide a foundation for set theory. [...] But clearly it is somehow circular to reduce the notion of set to a general notion of domain. (Skolem [1923], p. 292.)

The circularity would not obtain if we succeeded in describing a model without recourse to axiomatic set theory. Now the main thrust of Skolem's criticism of axiomatic set theory lies in the conviction that it *is* actually possible to produce such a particular model of (first-order) axiomatic set theory, but that such a model is only countable. The reason being that, for Skolem, uncountability is a concept internal to axiomatic set theory in the sense that there are no uncountable objects to be found by the means available outside or without axiomatic set theory.[39]

This statement deserves two notes. Firstly, one might object that countable and uncountable sets were first distinguished in the original set theory of Cantor's, which was not axiomatic. Cantor's informal set theory was not directly affected by the set-theoretic paradoxes, so it was not straightaway contradictory, as Skolem [1923], p. 291, claimed. Yet, we should not look to Cantor's set theory with much hope. On the one hand, it is crucial to ask what Cantor's proof of uncountability of all sets of integers relies on. The famous diagonal proof published in Cantor [1891] shows, in effect, that

famous debate concerning the notion of 'definit' property introduced by Zermelo [1908] in connection with his axiom of separation. The other participants in the debate were Weyl, Fraenkel, and Zermelo himself; this debate was one of the catalysts that contributed to the 'historical triumph' of first-order logic. For its history, see Shapiro [1991], pp. 181–190.

[39] Cf. Jané [2001], p. 143.

the sets of integers we are able to specify do not exhaust the totality of sets of integers, since we are always able to come up with a new set that has escaped our specification. In other words, for any particular specification, there always exist sets that have not been specified, and we can exemplify such sets by providing a specification for them. Assuming that the number of specifications we are able to provide is only countable,[40] the totality of specifiable sets is also only countable. Putting it all together, the key assumption behind the argument for the uncountability of the sets of integers is that there is a domain of sets of integers comprising the entirety of such sets, no matter whether specifiable or not. As Hallett [1984], p. 58, puts it, this assumption 'goes most of the way towards acceptance of the existence of *arbitrary* sub-domains', which, in the case of sets, may be viewed as an implicit assumption of the full power set operation. On the other hand, as I have already mentioned, if we want to prove that there are uncountable sets, the first step consisting in showing the uncountability of the sets of integers needs to be supplemented by the second step establishing that the sets of integers form a set. Unfortunately, as we pointed out in section 2.2, Cantor's concept of set was not sufficiently clear, and I do not wish to contribute to the debate of how to explicate it most naturally. Nevertheless, it seems uncontroversial to say that a collection of numbers—for instance, of the real numbers, $\mathbb{R}$—is a set provided it is a proper segment of numbers. This is nothing else than to say that it is a set only if there are greater numbers succeeding it.[41] Hence showing that the real numbers form a set requires establishing that there are greater collections than that of real numbers. The method for producing ever greater sets on the basis of those already given was, indeed, based on the diagonal argument introduced in Cantor [1891];[42] it was generalized in Cantor [1895], pp. 486–488, where it was called the 'exponentiation of powers'. The exponentiation provides the cardinal number of the set of all functions $f : \beta \mapsto \alpha$. In particular, let $\mathbb{N}$ be the set of natural numbers and $\aleph_0$ its cardinality. Then we obtain by exponentiation the cardinal number $2^{\aleph_0}$, which is the cardinality of the set of functions $f : \mathbb{N} \mapsto \{0, 1\}$. Although Cantor arguably never got as far as the standard power set operation,[43] his exponentiation shares with the former, again, the

[40] If we are strict and take into account the crude fact that our time is always limited, the number of specifications will be merely finite.

[41] An analogical idea is behind the guiding principle distinguishing sets from (proper) classes in the systems of set theory that include classes, such as von Neumann-Bernays-Gödel set theory (NBG) or Morse-Kelley set theory (MK). We will say more about these theories in section 3.5.

[42] Cf. Dauben [1979], p. 167.

[43] Lavine [1994], pp. 90–98, interprets Cantor's exponentiation of powers introduced in the 'Beiträge' as a form of the power set operation, and he sees in it a profound shift from Cantor's previous version of set theory developed in the *Grundlagen*. Lavine claims (p. 95) that, for the first time, Cantor was able to prove that the real numbers form a set, instead of merely assuming that they do. At the same time, the newly shaped set theory that

assumption that we are dealing with *all* the functions. So it can be concluded that Cantor's informal set theory employed—both with respect to proving the uncountability of the totality of all sets of natural numbers and with respect to arguing for the admissibility of such uncountable totalities as sets—virtually the same assumption Skolem argued against, namely that there is an uncountable totality of arbitrary functions.

Secondly, to say that there are no uncountable objects outside axiomatic set theory does not entail that the terms such as 'uncountable' cannot be used or understood outside set theory. After all, we saw in section 2.1 that it can be stated in Frege's system that there is no one-to-one correspondence between concepts or functions on the one hand and objects on the other, and we discussed in section 2.3 Russell's attempt to formulate a common principle behind all the set-theoretic paradoxes. Therefore, the concept of uncountability can be found and meaningfully employed outside axiomatic set theory.[44] (It is true that both systems, Frege's as well as Russell's, were strong enough to explicitly define the membership relation, so they incorporated a large degree of set theory. Yet, from this it only follows that some concepts of set theory are implicit in some presumably more basic and satisfactory concepts. These concepts are regarded to be of a logical character, and they are articulated as unprovable self-evident universal truths. This perspective characteristic of Frege and Russell is very different from that of Hilbert's—according to whom axioms are put forward as implicit definitions of whatever satisfies them—and does not require that we have set theory available in order to interpret our system.)

emerged in the 'Beiträge' made problematic some of the basic principles that had been so far taken for granted (cf. also Moore [1982], p. 46). This interpretation by Lavine was criticized, e.g., by Tait [2000], pp. 282–285, or Ferreirós [2007], p. 265. Ferreirós argues that the closest that Cantor comes to introducing the power set axiom is in the letter to Hilbert written on 10$^{\text{th}}$ October 1898, in which he shows that there is a one-to-one correspondence between the power set $\wp(\mathbb{N})$ and the set of functions from $\mathbb{N}$ to $\{0, 1\}$, and thus $\mathbb{R}$. In combination with two principles—namely that (i) whenever two collections can be put into one-to-one correspondence and one of them is an admissible set, the other is also an admissible set, and that (ii) the set $\wp(x)$ of all the subsets of a set x is an admissible set, which is a clear statement of the power set operation—this result enables Cantor to reach the desired conclusion that $\mathbb{R}$ form an admissible set. However, only two days later Cantor wrote in another letter to Hilbert that his argument supporting the principle (ii) was illusory. The acceptability of the power set operation was thus set in doubt. This leaves the whole argument that $\mathbb{R}$ form an admissible set inconclusive for Cantor.

[44]This perspective forces us to refuse, for instance, the distinction made by Resnik [1966], p. 435. He claims that the meaning of some of the terms of set theory, such as 'is a member of', is learned with the help of ordinary language; these terms are used and understood also outside axiomatic set theory. On the other hand, some other terms of set theory, e.g., 'uncountable' or 'limit ordinal', are learnt and used exclusively within set theory; learning these terms, Resnik asserts, equals learning set theory itself. On the basis of this distinction, Skolem's claim might be taken to be that the models of axiomatic set theory that can be described without the use of set-theoretic terms proper are always only countable. However, if we accept that the terms such as 'uncountable' can be meaningfully used outside set theory, this construal of Skolem's position can no longer be maintained.

Let us repeat once again that the key point of Skolem's critique is not that the concept of uncountability does not have a proper meaning outside axiomatic set theory but rather that uncountable *objects* cannot be shown to *exist* without relying on the machinery of axiomatic set theory. And unless we are able to show that there is an uncountable model without employing axiomatic set theory itself, we are trapped in a circle, assuming what we are striving to demonstrate.

3.5 Second-Order Set Theory

How does Zermelo's system fare with respect to the weighty scenery set up by "Skolemism"? To be sure, it does not suffice to respond to Skolem's critique merely by pointing out the second-order character of ZFC_2. Rather, the following question inevitably forces itself upon us: What is it that second-order variables of the language of ZFC_2 range over? We saw already in section 3.1 that Zermelo's system requires a standard semantics, i.e., the assumption that second-order variables range over *all* or *arbitrary* properties and relations (henceforth just 'relations') or functions, and not only those that are definable within the system. But we have to ask: what are these relations and functions? How are we to construe quantification over them? In standard second-order *logic*, the issue is relatively straightforward. First-order variables range over the members of the universe (individuals), while second-order variables range over (all or arbitrary) sets of individuals. However, in set theory, the members of the universe (other than the urelements) are called 'sets'; so already the first-order variables are supposed to range over all sets whatsoever, including the power sets of sets. This is to say that if we interpreted relations and functions as sets (of sets), our presumably second-order set theory would, as a matter of fact, collapse into a first-order system. And as we know, there is nothing in first-order systems articulated in countable languages that can assure us that we really quantify over all sets. How then should we deal with the second-order variables of ZFC_2?

It has become customary in set theory to follow Frege's impulse and replace the talk of relations and functions by the talk of their extensions, called 'classes'. A system of set theory with classes alongside sets was first developed in von Neumann [1925]. Owing to subsequent contributions by Bernays and Gödel, the resulting system is usually called 'von Neumann-Bernays-Gödel set theory' (NBG). However, NBG is a conservative extension of ZF, i.e., a sentence couched in the language of NBG without classes is a theorem of NBG if and only if it is a theorem of ZF. Classes are used mainly for expediency; they allow us, among other things, to replace axiom schemata with single axioms, i.e., they make room for finite axiomatizability.

The most prominent system of set theory with classes that is not conservative with respect to ZF is Morse-Kelley set theory (MK). The key step

that gives MK the extra strength over NBG is that it lifts the restriction forbidding us to substitute for '$\varphi(x)$' in the axiom schema of class comprehension:

$$\forall x \exists X(x \in X \leftrightarrow \varphi(x)), \tag{A-Cla}$$

where φ does not contain 'X' free, any formula containing quantification over classes. This is to say that, in MK, we are allowed to put for '$\varphi(x)$' any well-formed formula of the extended language whatsoever, including formulas obtained by the application of (A-Cla).[45] This gives rise to impredicative classes, i.e., classes defined by means of quantification over the totality of classes. The extended deductive power of MK over NBG and ZFC manifests itself in the fact that MK is capable of proving the consistency of these weaker theories.

It was proved by Weston [1977] that MK is 'almost the same theory' as ZF_2 (p. 499). More precisely, let φ be a formula of the language of MK and let φ^* be its translation into the language of ZF_2, in which each atomic part of the form $x \in X$ is replaced by one of the form $P(x)$. Then $\mathrm{MK} \vdash \varphi$ if and only if $\mathrm{ZF}_2 \vdash \varphi^*$. (To obtain this result, we also need to employ a logical principle of substitution for ZF_2 to the effect that $\forall P(\varphi(P)) \rightarrow \varphi(\psi(x))$, where '$\varphi(\psi(x))$' is the result of substituting a formula of the form $\psi(x)$ for '$P(x)$' (certain restrictions apply), and we need to devise a principle of identity for relations such as the one stating that $P = Q \leftrightarrow \forall x(P(x) \leftrightarrow Q(x))$.) This entails that MK and the second-order system of ZFC_2 are very similar in terms of their proof-theoretic capabilities.

However, their semantic properties significantly differ. The most important feature of MK in this respect is that it is not a genuine second-order theory. In fact, it is not second-order at all, and there is even a case to be made for considering it a single-sorted first-order theory. It is true that it distinguishes two kinds of objects, i.e., sets and classes, and that classes are intended to be the objects of the level $V_{\alpha+1}$, where V_α is the intended model of ZFC. However, MK habitually identifies sets with classes: every set is a class, and every class that has the same members as a set is identical with that set. The classes that are not identical with sets are called 'proper classes'. This identification allows us to explicitly define the property of being a set. Assuming there are no urelements, sethood can be defined as follows:

$$Set(X) \leftrightarrow_{Def.} \exists Y(X \in Y). \tag{D-Set}$$

This means, that in MK we could do with only a single sort of variables, namely the variables ranging over classes. In other words, two different sorts

[45] Contrary to NBG, MK is not finitely axiomatizable. NBG can be finitely axiomatized because the schema (A-Cla) is, in its restricted form, provably equivalent to the conjunction of a finite number of its instances, i.e., the axioms of class existence or class construction. Giving up the aforementioned restriction makes any finite number of instances insufficient to capture the full power of the class comprehension schema, thence the lack of finite axiomatizability of MK.

of variables are not used out of necessity, but rather for the sake of user-friendliness. A consequence of this slightly concealed first-order character of MK is that standard results concerning first-order theories apply. Among others, MK is affected by the Löwenheim-Skolem theorem, and has non-well-founded models. Moreover, the identification of sets with classes raises some important philosophical questions: What is it that prevents us from identifying all classes with sets?[46] If we cannot identify all classes with sets, how does it come about that we can identify any?[47] We will deal with some of the questions concerning the nature of classes in section 4.8.

If ZFC_2 is to be a genuinely second-order system, it has to be interpreted as involving two ultimately distinct kinds of objects. Sets must not be a mere subclass of classes but must form an autonomous domain. Yet, of the two orders of objects, sets are definitely the clearer ones; how should we think of properties and relations or functions? Given the relative obscurity of these notions when taken in themselves, we will interpret the second-level objects of ZFC_2 as classes. This decision is easily justifiable since properties can be represented by classes of objects, and relations and functions by classes of ordered n-tuples.[48] In any case, the penchant for classes in place of properties, relations and functions should be construed as a matter of convenience, a welcome simplification, rather than a decision with a decisive ontological impact. Nonetheless, if they are to be genuinely second-order, it is crucial to interpret classes as objects distinct from sets. In general, we will interpret classes as objects of exactly the level $V_{\alpha+1} = \wp(V_\alpha)$, where V_α is the domain of a model of ZFC_2. (Note that, when describing a model of ZFC_2, it is not necessary to specify a domain of classes over which the class variables range alongside that of sets. Once we have provided the set domain, we have implicitly given also the class domain: class variables range over all possible collections of the objects of the set domain.)

In this way, sets and classes remain separated as distinct kinds of entities even if they contain the same members. However, if a certain class contains the same objects as a particular set, there is nothing that prevents us from replacing the talk of that class by the talk of the corresponding set. The classes that do not have any counterparts among sets are proper classes. From this it follows that classes cannot reach out far from the hierarchy of sets in the sense that there cannot be a membership chain of exclusively proper classes such that $X_1 \in X_2 \in \ldots \in X_n$. The members of any membership chain of classes will always be, apart from the closing one, representable

[46] Cf. Tait [1998*a*], pp. 279–280.

[47] Cf. Potter [2004], pp. 313–315.

[48] This statement needs to be qualified: there are purposes for which the representation of relations and functions by means of classes is not viable, for instance, the purpose of dealing with so-called 'positive formulas' in the context of the higher-order reflection (cf. Tait [1998*a*], pp. 276, 283–289). Nonetheless, the representation of relations and function by classes suffices for everything we are going to do, so we will stick to it.

by sets. In other words, there is no hierarchy of proper classes alongside that of sets. Proper classes can be identified with the universe of all sets or with the collections that can be put into a one-to-one correspondence with the universe of all sets.[49]

3.6 Zermelo's Relativism

The universe of all sets V was defined on p. 55 as the union of all segments, $\bigcup_\alpha V_\alpha$. In an analogical manner, we can specify the totality Ω of all ordinal numbers as $\{x \mid x \text{ is a von Neumann ordinal}\}$. Yet, what are these specifications? What does it mean to say: 'all segments' or 'all von Neumann ordinals'? To specify the meaning of the quantifier binding the first-order variables is nothing else than to specify the domain of a model of the given theory; in the case of normal models, this amounts to specifying the domain V_α, where α is an inaccessible ordinal. As there is no normal model whose domain contains the universe of all sets in the absolute sense, the only definite meaning we are able to assign to the quantifier is restricted to the domain of this or that particular normal model. Hence, given any such normal model $\mathcal{M}$, V simply has to refer to the domain V_α of $\mathcal{M}$ and Ω has to refer to all the von Neumann ordinals within V_α. Then, indeed, there are just as many such totalities as there are models, and the meaning of an expression denoting such a totality becomes clear only once we have specified the model.[50] In other words, the symbols 'V' and 'Ω' as well as their specifications mentioned above do not denote totalities that would be preserved across models but their meaning changes from model to model.

The totalities such as those denoted by 'V' and 'Ω' are, indeed, proper classes. Recall now that the sequence of inaccessible ordinals, and hence also of normal models, is unbounded, and that, as we saw at the end of section 3.3, the domain of any normal model can become an object within the domain of another, higher model. From this it follows that proper classes of one model can become sets of another model. This means that both the notion of set and that of proper class become relative. It is important to understand the particular nature of this relativity. In a sense, the meaning of every symbol of any theory is relative with respect to a model since the model is precisely from where the meaning gets assigned. However, Zermelo's relativity is not of this crude sort. Rather, it consists in the claim that one theory can be so reinterpreted, i.e., provided with such a model, that the truth value of all the statements involving proper classes will be preserved, while the variables ranging over classes will be assigned values

[49] The statement that every subclass of the universe of sets is either a set or can be put into a one-to-one correspondence with the universe of sets is sometimes called 'von Neumann's principle'. It can be derived in NBG from class versions of the axiom of foundation, replacement and choice; viz. Smullyan and Fitting [1996], pp. 91–93.

[50] This argument can be found in Tait [1998*b*], p. 473.

from collections that are sets within a higher normal model.[51] If the axioms for sets are supposed to implicitly define what a set is, and if the axioms for classes are to implicitly define what a class is, then one has to conclude that, according to Zermelo's cumulative picture, no single system of set theory can claim to have achieved such a goal: what one system implicitly defines as a proper class will be a set in another system. Tait's saying: 'there are no absolute proper classes' (Tait [2005*a*], p. 142) may be appended by the rejoinder: 'no collection is so peculiar that it cannot become a set'.

This relativism of Zermelo's serves a double purpose. First, it provides the core of a solution to the set-theoretic paradoxes. Where Cantor distinguished between the determinate infinite totalities, i.e., sets, and the absolute or inconsistent infinite totalities, Zermelo supplied the distinction between (on the one hand) totalities that are so large that they cannot be captured *within* a given system of set theory and (on the other hand) collections that appear as ordinary sets in an extended system; the latter collections can, however, be shown within the extended system to be identical with the former, unsurmountable totalities. The distinction between the consistent and the inconsistent, which is arguably not very helpful,[52] has thus been supplanted by the distinction between the inside and the outside of a theory. This involves a significant shift in the way set theory is viewed and treated. We will return to the nature of this shift in a moment; for the time being, let us just say that it consists in Zermelo's replacement of Cantor's conception of a fixed or static universe of sets by the dynamic relativism of the essentially open cumulative hierarchy.

Secondly, and more importantly for us at the moment, the purpose of Zermelo's relativism is to make available a viable vindication of "genuinely uncountable" set theory against the dreaded dangers of "Skolemism". Writing about his allowing non-sets of one normal model to be genuine sets in a higher model, Zermelo says:

> By "relativizing" the notion of set in this way, I feel able to refute Skolem's "relativism" that would like to represent the whole of set theory in a *countable* model. It is simply *impossible* to give all sets in a constructive way [...] and any theory founded on this assumption would by no means be a theory of sets. [Zermelo's emphasis; quoted in Ebbinghaus [2007], p. 203.]

Given the relativism described above, the requirement that set theory should be a theory of *all* sets should also include all the non-sets, i.e., proper classes,

[51] We already saw this in section 3.3. Assume that we extend the original system of ZFC_2 by adding the axiom stating the existence of the first inaccessible ordinal, thus obtaining the system ZFC_2+I. By extensionality, the domain V_α of the model of ZFC_2 will be identical with an ordinary set in the domain V_β, $\beta > \alpha$, of the model of ZFC_2+I.

[52] Cf. Kolman [2008], p. 384.

since they can become sets provided that we assume higher and higher inaccessible ordinals. By such a reinterpretation of (the extensions of) proper classes of one system as sets of another system, set theory becomes capable of including also strongly impredicative collections among its objects. At the same time, the true object of set theory, namely the unbounded cumulative hierarchy of sets, cannot be exhausted by any single system since no system is powerful enough to provide us with all sets there are. From this perspective, it is manifest that confining our attention only to a hierarchy of countable sets is nothing but a mutilation: we deliberately cut off a large portion of healthy flesh and deal only with a part of the subject matter. Once again, what we come across here is the issue of generality. It is required that set theory should be a theory of *all* sets, and not merely of those manifesting certain favourable qualities.

Having covered the preliminaries, we can turn to the broader question we are pursuing, namely to Zermelo's response to the dangers of "Skolemism". As Zermelo's whole late conception of set theory is built on the grounds of second-order logic with standard semantics, which, as we know, amounts to nothing else than to assuming the existence of uncountable collections from the very beginning, Zermelo's defence against Skolem obviously cannot and does not lie in providing an independent justification for the existence of uncountable collections; the centre of Zermelo's response to Skolem lies elsewhere. We can put the whole issue slightly differently and say that the feature that makes ZFC_2 satisfiable only by models with uncountable domains is its ability to quantify not only over sets, i.e., "combinatorial" collections that can be reached by a repeated application of the operations sanctioned by the axioms and that can be located within a definite segment of the cumulative hierarchy, but also over arbitrary relations and functions (or, in our construal, classes). The latter entities represent collections that are, rather than by composition from elements, obtained by a "logical" division of the universe. In the simplest case of properties, the universe of sets is divided into two parts: one consisting of the sets that have the property in question, the other of those that do not. A pressing question, of course, is what these relations and functions (or classes) are, and how they are to be dealt with. Yet, what is even more disquieting is the very fact that it turns out to be necessary, for obtaining full set theory in the sense described above, to assume the existence of other kinds of collections besides sets. In other words, if we want to obtain Zermelo's quasi-categorical hierarchy of sets, and with it truly uncountable set theory, we have to go *beyond* sets. This is the "Skolemite" worry to which Zermelo's reply is directed.

Having explained all the things above, the reply does not come as anything new but is essentially a repetition. As Zermelo's conception involves a relativism of (in our construal) proper classes, according to which every proper class can eventually become a set, there is no requirement of going beyond sets in any absolute sense. Proper classes of one system of set theory

can become ordinary sets of an extended system, and so on. No stepping outside set theory is needed. Therefore, it is true that, in order to obtain a theory of truly uncountable sets, we have to quantify over non-sets but it is not true that we have to move beyond the cumulative hierarchy of sets. Once we do not restrict our view of set theory to a particular system of axioms, but rather consider it to be an open-ended sequence of stronger and stronger axiom systems that contain axioms asserting the existence of greater and greater inaccessible ordinals, some objects of set theory will become truly uncountable and set theory fully self-sustained, in the sense that no employment of any extra-set-theoretic assumptions is necessary.

There is a remarkable consequence of this view of set theory which is worth emphasizing. We already signalized on three separate occasions (discussing, first, Zermelo's refusal to accept the axiom of infinity as an axiom on a par with the other axioms of ZFC_2, secondly, the inclusion of urelements, and eventually, the insistence on comprehending everything that may, on any level, become a set) that set theory was conceived of by Zermelo so as to be as comprehensive as possible. It was meant to be a theory of unlimited generality in the sense that it was to concern whatever systems of objects satisfying the axioms. The act of narrowing down the variety of such systems by excluding urelements, by requiring them to contain infinite sets, or by letting go the collections that are not sets of ZFC_2 but only proper classes, and concentrating only on ZFC and the instrumental facets of set theory that are really needed for doing mathematics threaten to blur our understanding of what sets are as well as impair the philosophical significance of set theory as a whole. By contrast, the Zermelian view leads to a recognition of the inherently dynamic and open nature of the set-theoretic universe, which eludes any attempt at a definitive characterization by a fixed system of axioms. This dynamism is in a direct opposition not only to Cantor's conception of the fixed or static universe of sets,[53] but also to Hilbert's model-theoretic view of the axiom system as an implicit definition of a structure, or a number of structures, satisfying the axioms. The reason for the discrepancy does not lie in the fact that the axioms are incapable of characterizing the structure up to isomorphism but rather in a form of incompletability inherent in Zermelo's conception. As Hallett [1996], p. 1215, puts it, 'the very fixing of a model reveals an ordinal that cannot be in that model'. That is, the very determination of a structure satisfying the axioms immediately opens up objects that lie beyond that structure, forcing us to step onto a next layer of the hierarchy.[54] Set theory as a theory of all

[53]For the discussion of Zermelo's shift in conceiving of the universe of sets, see, e.g., Moore [1982], p. 271, Kanamori [2004], p. 528, or Ebbinghaus [2007], p. 194.

[54]For a discussion of differences between Zermelo's and Hilbert's views of formal theories and their models see Taylor [1993], pp. 542–545. In brief, Taylor argues that despite a shared method Zermelo's conception of the subject matter of set theory was not the model-theoretic one of Hilbert's.

sets, i.e., of the entire cumulative hierarchy, is inexhaustible by any single axiomatic system. The universe of sets can possibly only be exhausted by the universe of ever stronger set theories.

Chapter 4

Truth in the Hierarchy of Sets

In this chapter, we will elaborate upon some of the themes connected with the cumulative hierarchy developed by Zermelo. In particular, we will be interested in the possibilities of defining the property of truth and developing a theory of truth within the broadly Zermelian framework. Whereas the previous chapter mainly comprised the discussion of motives coming from Zermelo's groundbreaking paper, in this chapter we will set the historical concerns aside, and we will attempt to formulate some ideas concerning truth in a more or less systematic fashion. In the process of getting to the view concerning the property of truth we wish to defend, it will be necessary to introduce several auxiliary technical notions. They are really indispensable, so we have to undergo the pains and develop them in sufficient detail, even though they do not partake in our proper goals.

4.1 Arithmetization of Syntax

Truth or falsity are properties of sentences of a given language. A theory of sets such as ZFC_2 is formulated in the standard language of $\mathcal{L}_{\mathrm{ZFC}_2}$. The vocabulary of this language contains just the usual vocabulary of second-order logic plus a single two-place relation constant, '$\in$'. It does not contain any expressions that would make it possible to speak of sentences and other syntactic objects. As is well know, there is a technique that allows us to get past this difficulty by switching the talk of syntactic objects for the talk of the objects that a theory such as ZFC (or PA) can speak about, namely sets (or natural numbers). The technique of arithmetization was first developed in Gödel [1931], and since then it has become an indispensable tool in metamathematical investigations. Briefly, it consists in correlating the syntactic objects of a given language such as expressions, (well-formed) formulas, proofs etc.—in general, finite sequences of symbols—with natural

numbers or, if we are working in set theory, with sets. The correlation has to be workable in both directions: there has to be an encoding algorithm, i.e., a mechanical procedure which takes us in a finite number of steps from an expression of a given language to the natural number or the set associated with it, and there has to be a decoding algorithm which takes us back to the original expression. Indeed, the coding scheme can be designed in a wide variety of different ways. If it satisfies the aforementioned condition, it is deemed acceptable.

The goal of arithmetization of syntax is twofold. First, as already suggested, it enables us to convert assertions about syntactic objects into assertions about natural numbers or sets. Secondly, it makes it possible to convert some of these assertions about numbers or sets into sentences of the formal language.[1] Put together, the technique of arithmetization opens up the possibility for a formal theory to express and prove certain facts about numbers or sets that reflect its own syntactic properties. This is often expressed by means of autoreference: a system of arithmetization (or of 'Gödel numbering', as it is often called) is said to give a formal theory the ability to speak about itself. The talk of autoreference is fine provided we keep in mind that the sentences of a given formal theory do in reality speak only about numbers or sets—whose properties, however, reflect the properties of the correlated syntactic notions.

It is not necessary for our limited purposes to develop a coding scheme in full. Therefore, we will save ourselves the time and the effort, and we will merely sketch a simplified scheme in which expressions are represented by ordered n-tuples,[2] and we will do that just for a fragment of $\mathcal{L}_{\mathrm{ZFC}}$. That is, we will develop a coding scheme only for formulas containing just the "core" symbols of $\mathcal{L}_{\mathrm{ZFC}}$, namely the individual variables, the propositional connectives '$\vee$', '$\neg$', the relation expressions '$=$', '$\in$' and the existential quantifier '$\exists$'. It is well known that the remaining usual sentential connectives and the universal quantifier can be explicitly defined by means of these, hence any formula containing the additional connectives can be transformed into one that is composed merely out of the core ones. Alternatively, there is no obstacle to extending the coding scheme so that it covers also formulas containing the additional symbols.

The set correlated with a formula φ according to our coding scheme will be called the 'Gödel set' of φ, and it will be designated by '$\ulcorner\varphi\urcorner$'. Assume that we have a single infinite sequence of individual variables $v_1, v_2, \ldots, v_i, v_j, \ldots$, where any $i, j < \omega$. Then the Gödel sets are correlated with the formulas in the following way. For atomic formulas:

$$\ulcorner v_i = v_j \urcorner \text{ is } \langle 0, i, j\rangle,$$

[1] Cf. Enderton [2001], p. 224.

[2] This method is described in Drake [1974], p. 90. Drake attributes the original idea to Dana S. Scott but does not provide any reference.

$\ulcorner v_i \in v_j \urcorner$ is $\langle 1, i, j \rangle$.

For complex formulas:

$\ulcorner \varphi \vee \psi \urcorner$ is $\langle 2, \ulcorner \varphi \urcorner, \ulcorner \psi \urcorner \rangle$,
$\ulcorner \neg \varphi \urcorner$ is $\langle 3, \ulcorner \varphi \urcorner \rangle$,
$\ulcorner \exists v_i \varphi \urcorner$ is $\langle 4, i, \ulcorner \varphi \urcorner \rangle$.

Bearing in mind the acknowledged limitations, it may be maintained that our coding scheme provides any formula of $\mathcal{L}_{\mathrm{ZFC}}$ with its Gödel set, which is just an ordered n-tuple containing possibly other ordered n-tuples. Note that, as only formulas of finite length are accepted, every Gödel set belongs to the level V_ω of the cumulative hierarchy.

Later on, we will need to assume that we have a coding scheme also for the formulas of $\mathcal{L}_{\mathrm{ZFC}_2}$ as well as for more complex syntactic objects than formulas such as proofs. We will simply assume that we have such a scheme at our disposal, without bothering with describing it. To be sure, the scheme for $\mathcal{L}_{\mathrm{ZFC}_2}$ would need to incorporate the clauses for the higher-order variables. The coding for proofs, or sequences of formulas in general, would require a technique more sophisticated than that of the n-tuples described above. However, it is well known that all this can be done, and there is nothing particularly deep or exciting about it. Hence we can feel free not to stop here and carry on further.[3]

4.2 Truth for $\mathcal{L}_{\mathrm{ZFC}}$

Once the technique of arithmetization of syntax has been introduced, it becomes possible to work with the syntactic aspects of the given language and formally prove facts about them within a given axiomatic system. One might wonder whether the relations of satisfaction and truth for the given language are among those that are susceptible to such a treatment. After all, we have found a way of speaking about expressions, formulas or sentences within the given language, and the ability to express facts about sets was here from the beginning, so this objective might perhaps, at least at a first glance, seem feasible. Of course, it is well known that it is not. In this section we will examine why.

Let us just note that in the present section and the sections that follow, we will focus our attention primarily on the first-order system of ZFC rather than on second-order ZFC_2, with which we have been dealing for most of the time. The underlying language will be $\mathcal{L}_{\mathrm{ZFC}}$. This is because we want to make manifest that the definitions that will be given later can be formulated

[3] For a nice detailed treatment of the arithmetization of syntax, see, e.g., Smith [2007], pp. 124–137. Smith includes the proofs that some basic numerical relations reflecting the syntactic ones are primitive recursive.

in the first-order framework. Of course, as ZFC_2 is an extension of ZFC, it goes without saying that these definitions may also be given in ZFC_2; but since the second-order framework is not necessary, our prevailing focus will be confined to the first-order system of ZFC.

What do we require of the property of truth deserving that name? We will conform to the classic analysis given by Tarski in his monograph on truth[4] as it has become standard. If there is a set-theoretic property of truth, Tr, for $\mathcal{L}_{\mathrm{ZFC}}$, it should hold of (the Gödel set of) any $\mathcal{L}_{\mathrm{ZFC}}$ sentence φ if and only if φ is true. Now suppose that there is such a set-theoretic property, and that it is expressed by the formula $Tr(x)$ of an extended language $\mathcal{L}_{\mathrm{ZFC}+}$ which includes $\mathcal{L}_{\mathrm{ZFC}}$. The requirement imposed on the set-theoretic property of truth can then be formulated as the schema of $\mathcal{L}_{\mathrm{ZFC}+}$: $Tr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$. We may conclude that a formula is the set-theoretic TRUTH-PREDICATE for $\mathcal{L}_{\mathrm{ZFC}}$ if and only if this schema is true for every $\mathcal{L}_{\mathrm{ZFC}}$ sentence φ. In other words, 'Tr' is the set-theoretic truth-predicate for $\mathcal{L}_{\mathrm{ZFC}}$ if and only if its extension is precisely the set of the Gödel sets associated with the true sentences of $\mathcal{L}_{\mathrm{ZFC}}$.

It is one thing to be able to express by a predicate the property of truth that satisfies the aforementioned requirement, and another thing to be able to prove that the requirement is met. The former is a matter of the expressive richness of the language, the latter of the deductive strength of the theory. If a theory T is able to prove that the requirement is satisfied, i.e., if the following schema of $\mathcal{L}_{\mathrm{ZFC}+}$ (which is an extension of $\mathcal{L}_{\mathrm{ZFC}}$) holds for every sentence φ of $\mathcal{L}_{\mathrm{ZFC}}$:

$$\mathrm{T} \vdash Tr(\ulcorner\varphi\urcorner) \leftrightarrow \varphi, \tag{C-T}$$

we will say that T is the THEORY OF TRUTH or the DEFINITION OF TRUTH for $\mathcal{L}_{\mathrm{ZFC}}$. Of course, (C-T) generalized for any language and its extension is nothing else than the famous and widely debated convention T pushed forward by Tarski as the criterion of adequacy of any admissible truth definition.[5]

The question we asked at the outset of this section can now be sharpened. In fact, it is not one question but two: Is it possible to introduce in $\mathcal{L}_{\mathrm{ZFC}}$ the truth predicate expressing truth for $\mathcal{L}_{\mathrm{ZFC}}$? Is it possible to develop the theory of truth for $\mathcal{L}_{\mathrm{ZFC}}$ within ZFC? Similar questions may be asked about the relation of satisfaction; after all, as we will see, the concept of truth can be viewed merely as a particular representative of the more general concept of satisfaction. However, despite its being decisively more general, we will regard the relation of satisfaction as more or less auxiliary. This

[4] Cf. Tarski [1933], especially pp. 186–189.

[5] The convention T was first introduced in Tarski [1933], pp. 187–188. The biconditional part of (C-T) in its schematic form is nowadays usually called the 'T-schema', while the individual instances of the schema are called 'T-sentences'.

is why we will sometimes give preference in our mode of speaking to the concept of truth, bearing in mind, of course, that the two concepts are deeply intertwined.

We have suggested that the relations of satisfaction and truth for $\mathcal{L}_{\mathrm{ZFC}}$ cannot be defined in ZFC. Given what we have just said, this is to be read as claiming both that the set-theoretic property of truth for $\mathcal{L}_{\mathrm{ZFC}}$ cannot be expressed in $\mathcal{L}_{\mathrm{ZFC}}$, i.e., that there is no such truth predicate available in $\mathcal{L}_{\mathrm{ZFC}}$, and that ZFC does not contain the theory of truth for $\mathcal{L}_{\mathrm{ZFC}}$, i.e., that the set-theoretic property of truth for $\mathcal{L}_{\mathrm{ZFC}}$ cannot be explicitly defined in ZFC. It is rather important to indicate why the relations of satisfaction and truth cannot be expressed and defined in this way. Therefore, we will attempt to work out the "definitions" of these undefinable general relations in as much detail as possible. The effort will not be invested in vain since we will be able to use the modified versions of the definitions with better success later. To prevent confusion, we will mark the relations whose definitions are not acceptable by an asterisk.

As it has become our usual practice, we need to introduce a couple of auxiliary notions before getting to the thing itself. The first one is a three-place relation $Fm(u, s, n)$ obtaining between a Gödel set u associated with a formula, a formation function s and a natural number n.[6]

$$\begin{aligned} Fm(u,s,n) \leftrightarrow_{Def.}\ & \\ & Func(s) \wedge dom(s) = n+1 \wedge s(n) = u \wedge Int(n) \wedge \\ \forall k \leq n\big(\exists i,j < \omega[s(k) &= \langle 0,i,j\rangle \vee \\ s(k) &= \langle 1,i,j\rangle] \vee \\ \exists l,m < k[s(k) &= \langle 2,s(l),s(m)\rangle \vee \\ s(k) &= \langle 3,s(l)\rangle] \vee \\ \exists l < k \exists i < \omega[s(k) &= \langle 4,i,s(l)\rangle]\big). \end{aligned} \tag{D-Fm}$$

'$\exists i,j < \omega$' is, of course, the abbreviation for '$\exists i \exists j(i < \omega \wedge j < \omega \wedge \ldots)$'. '$Func(s)$', '$dom(s)$' and '$Int(n)$' mean '$s$ is a function', 'the domain of s' and 'n is a (non-negative) integer', respectively. These concepts are all definable in ZFC. In English, '$Fm(u, s, n)$' says that s is a function describing the formation of the set u as the Gödel set of a formula in $n + 1$ steps, starting from atomic formulas and continuing by induction. It can be used to obtain an explicit ZFC definition of the property of being a formula, namely: $Form(u) \leftrightarrow_{Def.} \exists n < \omega \exists s \in V_\omega(Fm(u, s, n))$. (The requirement imposed on s that it be finite is, of course, justified by the decision to accept only finite formulas in $\mathcal{L}_{\mathrm{ZFC}}$.)

To illustrate how the definition (D-Fm) works, let us describe an elementary example. Consider a rather simple formula: $\exists v_1 \exists v_2(v_1 \in v_2)$. This formula can be seen as constructed in three steps: (1) $v_1 \in v_2$, (2) $\exists v_2(v_1 \in v_2)$,

[6] The definition that follows is taken from Drake [1974], p. 91.

and finally (3) $\exists v_1 \exists v_2 (v_1 \in v_2)$. The construction of the resulting formula, φ, can thus be represented by means of a sequence $\varphi_0, \varphi_1, \varphi_2$, where φ_2 is φ. The purpose of the formation function s is to assign a particular value to each step; thus $s(0) = \langle 1,1,2\rangle$; $s(1) = \langle 4,2,\langle 1,1,2\rangle\rangle$; $s(2) = \langle 4,1,\langle 4,2,\langle 1,1,2\rangle\rangle\rangle$. The relation Fm then holds, in this particular example, if the Gödel set $\langle 4,1,\langle 4,2,\langle 1,1,2\rangle\rangle\rangle$ is the value that the function s assigns to the natural number 2, which it does. So the relation holds.

The remaining auxiliary relation that is required is the following three-place relation S^*:

$$\begin{aligned} &S^*(k,t,s) \leftrightarrow_{Def.} \\ &\exists i,j<\omega\big([s(k)=\langle 0,i,j\rangle \wedge t(k)=\{a|\ a(i)=a(j)\}]\vee \\ &\qquad [s(k)=\langle 1,i,j\rangle \wedge t(k)=\{a|\ a(i)\in a(j)\}]\big)\vee \\ &\exists l,m<k\big([s(k)=\langle 2,s(l),s(m)\rangle \wedge t(k)=t(l)\cup t(m)]\vee \\ &\qquad [s(k)=\langle 3,s(l)\rangle \wedge t(k)=\{a|\ a\notin t(l)\}]\big)\vee \\ &\exists i<\omega \exists l<k[s(k)=\langle 4,i,s(l)\rangle \wedge t(k)=\{a|\ \exists x(a(i/x)\in t(l))\}]. \end{aligned} \tag{D-S*}$$

Here '$a(i/x)$' signifies the function a with the value at i replaced by x, which can be defined in ZFC as $(a-\{\langle i,a(i)\rangle\})\cup\{\langle i,x\rangle\}$. The definition (D-S*) is fairly lengthy but the purpose of the relation $S^*(k,t,s)$ should be clear. Its goal is to correlate the formation function s, which assigns a "syntactic" value to every step k in the process of formation of the Gödel set of a formula, with the function t, whose value for any k is the set of appropriate assignments according to the particular value of $s(k)$. Thus S^* may be regarded as a function correlating sets that represent a "syntactic" facet of a formula with other sets that represent its "semantics".

In order to briefly illustrate how the definition works, let us take up again our previous example of the formula $\exists v_1 \exists v_2 (v_1 \in v_2)$. We have said that $s(2) = \langle 4,1,\langle 4,2,\langle 1,1,2\rangle\rangle\rangle$. Now to each step 0, 1 and 2 there corresponds a value of the function t. Thus we obtain the sequence $t(0) = \{a|\ a(1)\in a(2)\}$; $t(1) = \{a|\ \exists y(a(1)\in y)\}$; $t(2) = \{a|\ \exists x\exists y(x\in y)\}$. The relation S^* holds for $k=2$ if and only if $s(2) = \langle 4,1,\langle 4,2,\langle 1,1,2\rangle\rangle\rangle$ and $t(2) = \{a|\ \exists x\exists y(x\in y)\}$. Indeed, this is obviously so. So the relation S^* provides a correlation between $s(0)$ and $t(0)$, $s(1)$ and $t(1)$, etc.

With the auxiliary relations Fm and S^* in hand, let us try our luck and attempt to define the relation of satisfaction. This might seem to be doable by putting the relations Fm and S^* together and filling in some of the places occupied by the free variables:

$$\begin{aligned} &Sat^*(u,b) \leftrightarrow_{Def.} \\ &\qquad \exists t\exists s,n\in V_\omega\big(Fm(u,s,n)\wedge Func(t)\wedge \\ &\qquad dom(t)=n+1\ \wedge\ b\in t(n)\ \wedge\ \forall k\le n(S^*(k,t,s))\big) \end{aligned} \tag{D-Sat*}$$

In English, $Sat^*(u,b)$ says, with some simplification, that the Gödel set u associated with a formula is satisfied by the assignment b. To return to our

familiar example once again, the Gödel set $u = \langle 4, 2, \langle 1, 1, 2\rangle\rangle$ corresponding to the formula $\exists v_2(v_1 \in v_2)$ is satisfied by an assignment b if and only if $b \in \{a \mid \exists y(a(1) \in y)\}$, i.e., if $\exists y(b(1) \in y)$.

Finally, employing the relation Sat^* we can try to explicitly introduce the property of truth. For this, however, we would need a definition of the property of being the Gödel set associated with a *sentence* of $\mathcal{L}_{\mathrm{ZFC}}$. This property, $Sent(u)$, can be explicitly defined in a fashion similar to that of defining $Form(u)$ with the help of the formation function s. Alternatively, we could use the fact that the sentence is just a formula without free variables, and define a property that would hold of a Gödel set of a formula if and only if the associated formula did not contain any free variables. In any case, we will do ourselves the favour of simply assuming that we have the property $Sent$ at our disposal, without putting down its actual definition. Then we can state quite simply:

$$Tr^*(u) \leftrightarrow_{Def.} Sent(u) \wedge \forall b(Sat^*(u, b)), \qquad \text{(D-Tr}^*\text{)}$$

which says that the Gödel set u associated with a sentence is true if and only if it is satisfied by all assignments.

We have said that the relations of satisfaction and truth for $\mathcal{L}_{\mathrm{ZFC}}$ are not definable in ZFC. We have also duly marked the relations S^*, Sat^* and Tr^* with asterisks. What is wrong with these relations? Well, to put it harshly, it can be shown that the relations S^*, Sat^* and Tr^* do not exist. There does not exist a set whose members would be exactly the ordered pairs $\langle u, b\rangle$ such that u is the Gödel set of a $\mathcal{L}_{\mathrm{ZFC}}$ formula φ and φ is satisfied by the sets assigned to u by b. There does not exist a set whose members would be exactly the Gödel sets u associated with true $\mathcal{L}_{\mathrm{ZFC}}$ sentences φ. Of course, this is rather a rewording than an explanation, so we need to ask further: Why do these sets not exist?

The reason of the trouble lies already in the auxiliary relation S^*. It contains the abstraction operator '$\{ \mid \}$', and as we are restricted to the limited expressive resources of $\mathcal{L}_{\mathrm{ZFC}}$, we are bound to interpret it as a set. Thus if we have, e.g., $\{x \mid \varphi(x)\}$, we have to read it as $\exists z \forall x(x \in z \leftrightarrow \varphi(x))$ (z must not occur in φ). Take, for instance, the equation $t(k) = \{a \mid a(i) \in a(j)\}$ occurring in the second clause of (D-S*). It says that the value assigned to k by t is the *set* of the functions a such that the value a assigns to i is a member of the value a assigns to j. The domain of a is the set of sets representing the free variables of $\mathcal{L}_{\mathrm{ZFC}}$ formulas. Hence the domain of a is restricted to the lower levels of the cumulative hierarchy, and V_ω can be taken as their upper bound. Therefore, the claim that the domain of a is a set is fully justified. On the other hand, the range of a is not restricted at all. It comprises all sets whatsoever. Now any ordered pair $\langle x, y\rangle$, where x is a set representing a free variable and y any set in the universe, will be a set. However, the collection of all these ordered pairs will no longer be a

set, i.e., there is no set in the cumulative hierarchy that could be identified as $\{a \mid a(i) \in a(j)\}$. Assuming that such a set does exist quickly leads to a contradiction. This means that a property 'to be an assignment' does not have a set as its extension. So the definition (D-S*) has to be rejected as faulty, and the relations S^*, Sat^* and Tr^* must be dumped.

Is there any way that the definition (D-S*) could be rectified? Well, of course, probably everything can be rectified provided the changes go deep enough to tackle the root of the problem. In this particular case, it can be shown that no cosmetic changes suffice. The occurrence of the totality of all assignments in (D-S*) is not accidental. To be sure, it is certainly possible to put forward variants of the definition that proceed in different ways and shun the inclusion of the totality of all assignment functions. Nevertheless, unless such definitions differ in certain rather profound aspects that will be discussed in the sections to come, they are all bound to fail, too. The relations of satisfaction and truth are simply not representable as sets, and this claim can be supported by a general argument. A particularly simple version of it goes as follows.[7] Assume that we have a definition of truth for $\mathcal{L}_{\mathrm{ZFC}}$ and that 'Tr' is the truth predicate (in $\mathcal{L}_{\mathrm{ZFC}}$). Suppose that we have enumerated all formulas of $\mathcal{L}_{\mathrm{ZFC}}$ with one free variable, which gets us the sequence $\varphi_0, \varphi_1, \ldots, \varphi_n, \ldots$. Now let ψ be the formula $x \in \omega \wedge \neg Tr(\ulcorner \varphi_x(x) \urcorner)$. As ψ has one free variable, namely x, it will be among the formulas in the numbered sequence. Say that its number is k. Let us take k and replace with it the free variable in ψ; thus we obtain the sentence $\psi(k)$. Is this sentence Tr? We know that, according to the requirement imposed by the convention T, $Tr(\ulcorner \psi(k) \urcorner) \leftrightarrow \psi(k)$. However, $\psi(k)$ is by definition $\neg Tr(\ulcorner \varphi_k(k) \urcorner)$. As $\varphi_k = \psi$, we get that $\psi(k)$ is $\neg Tr(\ulcorner \psi(k) \urcorner)$. Yet, by substituting back into the convention T condition we obtain $Tr(\ulcorner \psi(k) \urcorner) \leftrightarrow \neg Tr(\ulcorner \psi(k) \urcorner)$, which is a contradiction. So we may conclude that no matter how Tr is defined, the mere assumption that there is a truth predicate in the given language $\mathcal{L}$ that expresses truth for $\mathcal{L}$ leads to a contradiction.

Specifically, if we turn to ZFC, it may be asserted that ZFC cannot define truth for $\mathcal{L}_{\mathrm{ZFC}}$ and there is no predicate of $\mathcal{L}_{\mathrm{ZFC}}$ that expresses the set-theoretic property of truth for $\mathcal{L}_{\mathrm{ZFC}}$. In its general form, this result is known as Tarski's theorem.[8] Now what moral should we take from the inexpressibility and undefinability of the relations of satisfaction and truth within one and the same language? Does it mean that we have to give up all hope and stay clear of these relations for good? In what follows we will show how partial truth definitions can be given within ZFC and for $\mathcal{L}_{\mathrm{ZFC}}$ in

[7] This version of the argument is inspired by the one be found in Jech [2003], p. 162.

[8] The original formulation can be found in Tarski [1933], p. 247, and in the postscript, p. 273. Despite the fact that the theorem bears Tarski's name, there is some evidence that it was at around the same time independently discovered by Gödel and perhaps a little later by Carnap. Cf. Gödel's Princeton lectures, Gödel [1934], pp. 63–65, and Carnap [1934*b*], pp. 207–222.

certain extended systems. However, before we can plunge into that task we will need to absorb some more technical buildup.

4.3 The Hierarchy of Formulas

We have spoken of second-order set theory, second-order variables, etc. What are these higher-order objects? How are the formulas containing the variables ranging over the higher-order objects to be treated? As a first step towards answering some of these questions, let us introduce a system of finite types.[9] The FINITE TYPES are defined inductively: let $n \in \omega$ be ≥ 0 and let $\tau_1, \ldots, \tau_n$ be finite types; then $\tau = (\tau_1, \ldots, \tau_n)$ is a finite type. If $n = 0$, the type $\tau = ()$ is the type of sets. The type $\tau = ((), ())$ is the type of binary relations between sets and so on. It is useful to introduce, alongside the concept of type of a set, also the order of (the type of) a set. The ORDER of the type is a natural number obtained as follows: the order of $\tau = ()$ is 1; the order of $\tau = (\tau_1, \ldots, \tau_n)$ is 1 greater than the maximum order of $\tau_1, \ldots, \tau_n$. Note that the order is not a brand new concept that would bifurcate the classification of sets into types and orders, like in Russell's ramified hierarchy; the order is rather an abstraction from the concept of type neglecting other compositional features of objects that the type reflects. In the particular case of our class construal of ZFC_2, we are dealing with objects of types $\tau = ()$ of order 1 and $\tau = (())$ of order 2. The reason that we can do without the more complex types of second-order relations is, as we have already said, that they can be represented by classes.

To different types and orders of objects there correspond different types and orders of variables. This makes it possible to speak of the order of a formula of the given language, which paves the way for the introduction of a whole classification of formulas in terms of their order and quantificational complexity. In general, the position of a formula within the hierarchy is determined by its unbounded quantifiers.[10] For reasons to be explained shortly, we count neither the occurrence of free variables nor of the bounded quantifiers as contributing to the resulting classification of the formula. The hierarchy can be described as follows. $\Pi^0_0 = \Sigma^0_0$ is the class of formulas without unbounded quantifiers. It is customary to take this class to be of order 0. Σ^n_{m+1} is the class of formulas of the form $\exists X \varphi$ such that φ is Π^n_m and the order of the variable X is $n + 1$. Analogically, Π^n_{m+1} is the class of formulas of the form $\forall X \varphi$ such that φ is Σ^n_m and the order of the variables X is $n + 1$. (The reason why the order of X is to be $n + 1$ is historical: it would be more natural to have the order of the variables simply n instead of

[9] The system of finite types sketched here is developed in Kreisel and Krivine [1966], pp. 90–95, or Tait [1998*a*], pp. 275–276, Tait [2005*a*], pp. 142–144, and van Benthem and Doets [1983], pp. 306–308.

[10] A quantifier is bounded if it is of the form $\forall x(x \in y \to \ldots)$ or $\exists x(x \in y \wedge \ldots)$, which we will abbreviate as $\forall x \in y(\ldots)$ or $\exists x \in y(\ldots)$, respectively.

$n+1$ but the classification is well established in this shape, with the upper index '0' for first-order formulas, '1' for second-order formulas etc., so we will stick to the tradition.) Note that it is only quantifiers of variables of the order $n+1$ that are considered as contributing to the complexity of a Π^n_m or Σ^n_m formula.

Before continuing with the presentation of the hierarchy, we had rather pause and clarify several points. The hierarchy of formulas just described has it origins in the 1940s in Kleene [1943] and Mostowski [1947], and it has become standard for dealing with the complexity of formulas and definability of relations (about which we will speak shortly). However, it is often presented differently. To understand that the differences are usually not relevant, we need to bring in some of the basic facts related to it. Firstly, the Π^n_m or Σ^n_m formulas are often defined as those having not m single alternating quantifiers but rather m blocks of like quantifiers, followed again by a bounded or quantifier-free kernel. In fact, provided that we have a suitable form of induction—which is fulfilled both in Peano arithmetic, which contains the induction schema, and in ZFC, in which we can use the axiom of pairing—it can be proved that a formula with m alternating blocks of like quantifiers is equivalent to a formula with m single alternating quantifiers.[11] That is, the blocks of like quantifiers can be contracted.

Secondly, why have we discarded free variables and the bounded quantifiers as far as the complexity of a formula goes? The reason for this disregard is that a mere occurrence of a variable of order n in a formula does not always necessitate accepting the existence of a well delimited domain of objects of order n. It is the quantification that forces us to include the n-th order domain in the interpretation of the given language, and to take the higher-order objects seriously.[12] As regards the bounded quantifiers, we will say the following. In a theory such as Peano arithmetic this is straightforward as a bounded quantifier ranges only over a finite number of objects, so it can be replaced by a finite conjunction or disjunction. In standard set theory the use of a bounded quantifier entails that we do not have to check the whole hierarchy but we can stop at a level of a specific rank. Syntactically, the axiom of replacement makes it possible to move all the bounded quantifiers to the right and unbounded to the left,[13] so there is a clear cut between the bounded kernel, the range of whose bound variables is restricted to a specific V_α, and the rest of the quantifiers ranging over the whole universe. Thus, despite the fact that the given α may be infinite so there is no longer the possibility to get rid of the bounded quantifiers completely in favour of finite conjunctions or disjunctions, the analogy with arithmetic is preserved in the sense that, in order to check the given formula, we need to check the

[11] For a partial proof viz. Smith [2007], p. 75.
[12] Cf., for example, van Benthem and Doets [1983], p. 309, on this subject.
[13] See, for instance, Kanamori [2003], p. 5, to see how this is done.

sets only up to the αth level of the hierarchy. It is for a similar reason that we disregard in our classification the quantifiers of variables of order $< n+1$. In the presence of these variables we need to check only up to the types of a specific order $< n+1$ within the hierarchy of types whereas for checking the higher-order bits of the formula we have to go further.

Now let us complete the presentation of the hierarchy. We will say that a formula is Π^n_m or Σ^n_m even if it does not have the required form but is merely provably equivalent to a Π^n_m or Σ^n_m formula. This concession, however, requires a further specification. As any formula is equivalent to all formulas obtained merely by prefixing "vacuous" quantifiers that do not contribute to the contents of the formula in any significant way, it is useful to consider a formula to be Π^n_m or Σ^n_m where m is the smallest possible. The extension of the classification to equivalent formulas opens room for the following general result. It is a theorem of higher-order logic that every formula of order $n+1$ is provably equivalent to one in the class $\bigcup_{m<\omega}(\Pi^n_m \cup \Sigma^n_m)$, i.e., that every formula of order $n+1$ is equivalent to some Π^n_m formula or some Σ^n_m formula.[14] Hence our classification exhausts the totality of formulas. Finally, let us add that a formula is Δ^n_m if it is in the class $\Pi^n_m \cap \Sigma^n_m$, i.e., if it is equivalent both to a Π^n_m formula and a Σ^n_m formula.

The classification of formulas just developed can be used for a classification of definable relations. Let $\mathcal{M}$ be a model with the domain A, let R be a relation of type $\tau = (\tau_1, \ldots, \tau_n)$ over A, let $S_1, \ldots, S_n$ be objects of types $\tau_1, \ldots, \tau_n$ over A, and let φ be a formula of a given language with $X_1, \ldots, X_n$ free variables of types $\tau_1, \ldots, \tau_n$. Then we say that the formula φ DEFINES the relation R on $\mathcal{M}$ if

$$R(S_1, \ldots, S_n) \text{ if and only if } \mathcal{M} \models \varphi[S_1, \ldots, S_n], \qquad \text{(D-Def)}$$

where '$[S_1, \ldots, S_n]$' signifies an assignment of values $S_1, \ldots, S_n$ for the free variables $X_1, \ldots, X_n$. Employing the concept of definability, the classification of formulas may now be transformed into one for relations. A relation belongs to the class Π^n_m (Σ^n_m) if it is definable by a Π^n_m formula (a Σ^n_m formula). A relation is said to be in the class Δ^n_m if it is in $\Pi^n_m \cap \Sigma^n_m$, i.e., if it is definable both by a Π^n_m formula and a Σ^n_m formula. It is worth mentioning two particular applications of this general classification of relations in terms of definability. If we restrict our attention to the structure $\mathcal{N} = \langle \mathbb{N}, +, \times, 0, s \rangle$, the hierarchy of Π^0_m and Σ^0_m relations on $\mathcal{N}$ is called the 'arithmetical hierarchy'. The hierarchy of Π^1_m and Σ^1_m relations on $\mathcal{N}$ is called the 'analytic hierarchy'.

We are, nevertheless, interested in classifying the formulas of the language of set theory. We will say that a formula φ is Π^{ZFC}_m (Σ^{ZFC}_m) if it is provable in ZFC that φ is equivalent to a Π^0_m (Σ^0_m) formula ψ, i.e., if $\mathrm{ZFC} \vdash \varphi \leftrightarrow \psi$. Similarly, a formula is $\Pi^{\mathrm{ZFC}_2}_m$ ($\Sigma^{\mathrm{ZFC}_2}_m$) if it is provably

[14] For the proof, see van Benthem and Doets [1983], pp. 309–310.

equivalent in ZFC_2 to a Π^1_m (Σ^1_m) formula. Relations are classified in a similar manner. This hierarchy, which may, indeed, be generalized to apply to any particular theory T, is usually called the 'Lévy hierarchy'.[15] It is often extended to cover also terms; a term t is Π^{T}_m (Σ^{T}_m) if the formula $x = t$, where x does not occur within t, is Π^{T}_m (Σ^{T}_m). The inclusion of terms makes it possible to classify besides definable relations also functions. It remains to append that we will call the particular value of the lower index m the 'degree' of a formula. It is quite common also to speak of the 'rank' of a formula in this respect but in order to prevent any mix-up with the notion of the rank of a set, I have decided to opt for the unambiguous 'degree'.[16]

Adopting the Lévy hierarchy leads to certain important results concerning satisfiability of various kinds of formulas. The question can be raised: if a formula is satisfied by a particular subdomain of the universe V, is it also satisfied by the whole V? And vice versa: if a formula is satisfied by the whole universe V, is it also satisfied by a particular subdomain of V? Recall that any segment V_α is transitive. Let $\langle M, \in\rangle$ be any transitive substructure of $\langle V, \in\rangle$, and let $\varphi(x_1, \ldots, x_n)$ be any formula of the given language. Now if the following holds for any $a_1, \ldots, a_n \in M$:

$$\langle M, \in\rangle \models \varphi(a_1, \ldots, a_n) \leftrightarrow \langle V, \in\rangle \models \varphi(a_1, \ldots, a_n), \qquad \text{(D-Abs)}$$

we will say that φ is ABSOLUTE for M. If we have only the implication from the left to the right, i.e., $\langle M, \in\rangle \models \varphi(a_1, \ldots, a_n) \rightarrow \langle V, \in\rangle \models \varphi(a_1, \ldots, a_n)$ for any $a_1, \ldots, a_n \in M$, φ is said to be UPWARD PERSISTENT for M. If just the opposite direction holds, namely $\langle V, \in\rangle \models \varphi(a_1, \ldots, a_n) \rightarrow \langle M, \in\rangle \models \varphi(a_1, \ldots, a_n)$ for any $a_1, \ldots, a_n \in M$, φ is called DOWNWARD PERSISTENT for M.

Suppose that M is a transitive subdomain of V. Then it can be proved that the following kinds of formulas have the following properties:[17]

$\Pi^{\mathrm{ZFC}}_0 = \Sigma^{\mathrm{ZFC}}_0 = \Delta^{\mathrm{ZFC}}_0$ formulas are absolute for M;
Σ^{ZFC}_1 formulas are upward persistent for M;
Π^{ZFC}_1 formulas are downward persistent for M;
$\Pi^{\mathrm{ZFC}}_1 \cap \Sigma^{\mathrm{ZFC}}_1 = \Delta^{\mathrm{ZFC}}_1$ formulas are absolute for M.

Thus, for transitive structures, bounded quantifiers preserve absoluteness, while existential quantifiers preserve upward persistence and universal quantifiers preserve downward persistence. (The concept of absoluteness will not

[15] After Azriel Lévy who announced it in Lévy [1959] and published in full in Lévy [1965], pp. 4–11. For a full development of the Lévy hierarchy and a classification of some basic formulas and relations definable in ZFC with detailed proofs, see Drake [1974], pp. 76–89, or Devlin [1984], pp. 24–31.

[16] Lévy [1965], for instance, uses 'rank'. On the other hand, 'degree' has been used in a variety of different meanings; viz., e.g., Smullyan and Fitting [1996], p. 130. The terminology does not seem to be settled.

[17] For the proof see Devlin [1984], pp. 27–28.

be employed until section 4.7 but the most convenient way was to introduce it here with the other auxiliary notions.)

This terminates our technical buildup. We have absorbed enough for the time being, and we may get back to the relations of satisfaction and truth we left in section 4.2.

4.4 Truth in a Set

Having suffered a failure with our first attempt at a definition of truth for $\mathcal{L}_{\mathrm{ZFC}}$, it is clear that we have to change the strategy. We clearly understand that to provide a full definition of truth for $\mathcal{L}_{\mathrm{ZFC}}$ is not possible. Nevertheless, we can attempt to restrict the scope of the definition, and introduce only partial relations of satisfaction and truth. There are two broad lines how this idea might be developed. Either we can restrict the domain of the relation of satisfaction, i.e., the totality of objects which are assigned as values to the free variables occurring in formulas, or we can make use of the hierarchy of formulas introduced in the preceding section and define such a relation of satisfaction that will be applicable only to formulas of lower complexity. Both paths are walkable.

Let us start with the former. It turns out that it is possible to define a restricted property of truth for $\mathcal{L}_{\mathrm{ZFC}}$, namely the truth in a set. The idea is, as might be expected, that if we take a particular set, we will be able to define the set of sentences that are true in that set. Obviously, this restricted concept of truth will never be large enough to coincide with the totality of all the true sentences of $\mathcal{L}_{\mathrm{ZFC}}$ but we know that we must not hope to achieve that.

Fortunately, most of the work carried out in section 4.2 can be reused without much change. The definition (D-Fm) of the relation Fm, the only one without the asterisk, can be taken as it stands. The definition (D-S*) of the auxiliary relation S^* has to be modified by incorporating certain restrictions on the assignment functions a. The domain of a is the set of the sets representing the free variables occurring in formulas of $\mathcal{L}_{\mathrm{ZFC}}$. The supply of variables is infinite but as we admit only finite formulas, the actual number of the sets that need to be assigned values by a will always be finite, and cannot exceed that of the cardinality of the Gödel set u of a given formula φ. Therefore, we can take the rank $r = \rho(u)$ (i.e., $r = \rho(\ulcorner\varphi\urcorner)$) as the convenient upper bound for the domain of a. How should the range of a be restricted? We will permit the range of a to be any *set*; it will be indicated by the free variable w. Thus the assignment functions a will be functions $a : r \mapsto w$. To designate the set $\{f \mid f : r \mapsto w\}$, we will follow Drake [1974], p. 81, and use the expression '$^r w$'. (Note that this set is Δ_1^{ZFC}.) This restriction guarantees that the range of the function t will be a set, and that t itself will be representable as a set. If we include all these

modifications, the relation S_s, a replacement of the faulty relation S^*, may be defined as follows:[18]

$$\begin{aligned} S_s(k,t,s,r,w) \leftrightarrow_{Def.}\ & \\ \exists i,j<\omega\big([s(k)=\langle 0,i,j\rangle \wedge t(k)=\{a\in {}^r w|\ a(i)=a(j)\}]\vee & \\ [s(k)=\langle 1,i,j\rangle \wedge t(k)=\{a\in {}^r w|\ a(i)\in a(j)\}]\big)\vee & \\ \exists l,m<k\big([s(k)=\langle 2,s(l),s(m)\rangle \wedge t(k)=t(l)\cup t(m)]\vee & \\ [s(k)=\langle 3,s(l)\rangle \wedge t(k)={}^r w-t(l)]\big)\vee & \\ \exists i<\omega\exists l<k[s(k)=\langle 4,i,s(l)\rangle \wedge t(k)=\{a\in {}^r w|\ \exists x\in w(a(i/x)\in t(l))\}]. & \end{aligned} \tag{D-S$_s$}$$

(The lower index '$_s$' is used to indicate that the relation S_s is only defined for sets.) We already understand the purpose of this relation as well as the mechanism of the definition, so no further explanations are needed.

The relation of satisfaction Sat_s can now be defined in very much the same way as the faulty relation Sat^*, except for the need to bind the additional variable r:

$$\begin{aligned} Sat_s(u,w,b) \leftrightarrow_{Def.}\ & \\ & \exists t\exists s,n,r\in V_\omega\big(Fm(u,s,n)\wedge r=\rho(u)\wedge Func(t)\wedge \\ & dom(t)=n+1\ \wedge\ b\in t(n)\ \wedge\ \forall k\le n(S_s(k,t,s,r,w))\big) \end{aligned} \tag{D-Sat$_s$}$$

As expected, '$Sat_s(u,w,b)$' can be read as saying that the Gödel set u (associated with a formula of $\mathcal{L}_{\mathrm{ZFC}}$) is satisfied in the set w under the assignment b. Observe that the formula on the right-hand side of the definition is a Σ_1^{ZFC} formula. However, it is provable in ZFC that this formula is equivalent to an alternative defining formula which is Π_1^{ZFC}.[19] This makes the relation Sat_s Δ_1^{ZFC}.

Having amended the definition of Sat_s, it remains only to replace the definition (D-Tr*) by the following one:

$$Tr_s(u,w) \leftrightarrow_{Def.} Sent(u)\wedge \forall b\in {}^{\rho(u)}w(Sat_s(u,w,b)), \tag{D-Tr$_s$}$$

which says that the Gödel set u (associated with a sentence of $\mathcal{L}_{\mathrm{ZFC}}$) is true in w, i.e., satisfied by all assignments $b:\rho(u)\mapsto w$. This truth relation is also Δ_1^{ZFC}.

We will now introduce a handy notational convention. Let u be the Gödel set of a $\mathcal{L}_{\mathrm{ZFC}}$ formula φ. We will sometimes use, instead of '$Sat_s(u,w,b)$', the fancier notation '$w\models_s \ulcorner\varphi\urcorner\,[b]$'. And if u is the Gödel set of a $\mathcal{L}_{\mathrm{ZFC}}$ sentence φ, we will feel free to replace '$Tr_s(u,w)$' with '$w\models_s \ulcorner\varphi\urcorner$'.

It is crucial to understand clearly what relations we have just defined. The relations Sat_s and Tr_s have been defined solely for sets. This means

[18] This definition as well as the next one can be found in Drake [1974], p. 91.

[19] See Drake [1974], pp. 91–92. The equivalent formula is: $\exists s,n,r\in V_\omega\big(Fm(u,s,n)\ \wedge r=\rho(u)\wedge\forall t[(Func(t)\wedge dom(t)=n+1\wedge\forall k\le n(S_s(k,t,s,r,w)))\to b\in t(n)]\big)$.

that one can, for example, easily obtain the set x of (the Gödel sets of) the sentences of $\mathcal{L}_{\mathrm{ZFC}}$ that are true in $\mathcal{M}_\omega$ as $x = \{y \mid Tr_s(y, V_\omega)\}$. We know that $\mathcal{M}_\omega$ is a model of ZFC without the axiom of infinity; thus we are able to define in ZFC a particular set of (the Gödel sets of) those $\mathcal{L}_{\mathrm{ZFC}}$ sentences that are true in the set of finite sets. One more example: $\mathcal{M}_{\omega+\omega}$ is, among others, a model of ZFC without the axiom of replacement. Again, we can take the set $V_{\omega+\omega}$ and define—with the help of the relation Tr_s—the set of all the $\mathcal{L}_{\mathrm{ZFC}}$ sentences that are true in $V_{\omega+\omega}$. In general, the relation Tr_s makes it possible to pick out any set whatsoever and define the set of all the sentences of $\mathcal{L}_{\mathrm{ZFC}}$ that are true in that set. The trouble is, of course, that this relation is inadequate if what we are after is the general concept of set-theoretic truth, namely the set of (the Gödel sets of) sentences that are true in the set-theoretic universe, i.e., in $\mathcal{M} = \langle V, \in_V \rangle$ since the set of (the Gödel sets of) the sentences that are Tr_s can never be made sufficiently exhaustive. Moreover, not only that our definition cannot provide the general account of truth in the set-theoretic universe, it does not even suffice to provide a set of truths large enough to contain all (the Gödel sets of) the theorems of ZFC. For this, as we know, we would have to take the set V_α with α an inaccessible ordinal, whose existence is unprovable in ZFC but would have been required for the definitions of Sat_s and Tr_s to work. Thus the power of the partial relations Sat_s and Tr_s is really very much limited.

4.5 Truth in a Large Set

So Tr_s gives us only a fragment of set-theoretic truth. Does it mean that we cannot get more out of the strategy of restricting the domain of the satisfaction relation? No, fortunately not. This strategy can be successfully combined with the idea of enlarging the theory by adding new axioms. In this way, we will be able to reach the point when the fragment of the $\mathcal{L}_{\mathrm{ZFC}}$ sentences that are Tr_s in a suitably chosen set becomes identical with the totality of truths in the model of the original theory. Recall the picture drawn by Zermelo we discussed in sections 3.4 and 3.6, according to which no single system of set theory can exhaust the totality of sets, and, at the same time, there is no domain that cannot be reinterpreted as a set.

As we know, if one desires to transform the domain of a model of ZFC or ZFC_2 into a set, it suffices to add an axiom asserting the existence of the requisite inaccessible ordinal indexing the particular level of the cumulative hierarchy. There is a variety of ways of formulating such an axiom, hinging on the question whether we merely wish to assert that there is an arbitrary strongly inaccessible ordinal or that there is some particular such ordinal, e.g., the smallest one, or whether we aim at asserting the existence of a whole transfinite sequence of inaccessible ordinals, as in the following case:

$$\forall\alpha\exists x(Inac(x) \wedge \alpha < x), \qquad \text{(A-Ina)}$$

where '$Inac(x)$' means 'x is a strongly inaccessible ordinal'.[20] The axiom (A-Ina) asserts that for every ordinal α there is an inaccessible ordinal greater than α. It immediately gives rise to the transfinite sequence $\theta_1 < \theta_2 < \ldots < \theta_\alpha < \ldots$ of inaccessible ordinals, which may be viewed as materializing the conclusion of Zermelo's informal argument in favour of the existence of the unbounded sequence of ever greater inaccessible ordinals. It does not really matter for our purposes which of the various axioms asserting the existence of inaccessible ordinals we opt for. The decisive feature they all have in common is that once adopted, i.e., once ZFC has been extended into ZFC + a version of the axiom of inaccessibles, it becomes possible to define truth in a set corresponding to the domain of the model of the original, unextended ZFC. Let us choose ZFC+A-Ina, i.e., ZFC extended by the full power of the axiom (A-Ina). In this theory, the levels $V_{\theta_1}, V_{\theta_2}, \ldots, V_{\theta_\alpha}, \ldots$ evidently become ordinary sets within the set-theoretic hierarchy, whose existence is trivially provable. This gives us a repository of sets that can be employed in defining truth for ZFC.

At this point, however, we face the following problem. Both ZFC and ZFC+A-Ina share the same language, i.e., $\mathcal{L}_{\mathrm{ZFC}}$. But we need some means of distinguishing between the sentences of $\mathcal{L}_{\mathrm{ZFC}}$ that can be shown to be true in any model satisfying ZFC, on the one hand, and the sentences that are true in a model of ZFC+A-Ina, on the other. This is to say that we need something amounting to a restriction circumventing the former sentences within the latter. To get this restriction, let us introduce the well-known technique of relativization. The RELATIVIZATION φ^s of a formula φ to a set s is defined inductively:

if φ is atomic, i.e., either $x = y$ or $x \in y$, φ^s is φ,
$(\varphi \vee \psi)^s$ is $\varphi^s \vee \psi^s$,
$(\neg\varphi)^s$ is $\neg(\varphi^s)$,
$(\exists x\varphi)^s$ is $\exists x \in s(\varphi^s)$, i.e., $\exists x(x \in s \wedge \varphi^s)$.

To put it briefly, the relativization φ^s makes all the quantifiers occurring in φ bounded in s. (Note that if φ already contains bounded quantifiers, say that φ is relativized to t, φ^s entails a further restriction: $(\varphi^t)^s$ is equivalent to $\varphi^{(t \cap s)}$.[21])

Now let θ_1 be the smallest inaccessible ordinal, whose existence is provable in ZFC+A-Ina, and let $\varphi^{V_{\theta_1}}$ be a formula of $\mathcal{L}_{\mathrm{ZFC}}$ relativized to the level V_{θ_1} of the cumulative hierarchy. Then the definition of the relation $Sat_s(u, w, b)$ is applicable, and the formula

$$V_{\theta_1} \models_s \ulcorner \varphi^{V_{\theta_1}} \urcorner [b]$$

[20] This axiom, with the difference that there it involves inaccessible cardinals instead of ordinals, is stated in Drake [1974], p. 68. Drake attributes it to Tarski. Kanamori [2003], pp. 20–21, confirms the attribution, making reference to two articles of Tarski's published at the end of the 1930s.

[21] Cf. Drake [1974], p. 99.

defines the relation of satisfaction in V_{θ_1}, obtaining between (the Gödel set of) a formula $\varphi^{V_{\theta_1}}$ and the assignment b. Likewise for $Tr_s(u, w)$ and truth: if φ is a sentence, the formula

$$V_{\theta_1} \models_s \ulcorner \varphi^{V_{\theta_1}} \urcorner$$

defines the property of being true in V_{θ_1}, holding of (the Gödel set of) a sentence $\varphi^{V_{\theta_1}}$. Obviously, the formula immediately yields the set of the relativized "truths-in-V_{θ_1}", namely the set $x = \{\ulcorner \varphi^{V_{\theta_1}} \urcorner \mid V_{\theta_1} \models_s \ulcorner \varphi^{V_{\theta_1}} \urcorner\}$.

It is very easy to see that there is no longer an immediate danger of contradiction that would illegitimize the existence of this set of (the Gödel sets of) truths. Recall the argument against the definability of truth on p. 86. The pivotal role in the various arguments against the definability of truth is played by the idea of diagonalization: typically a formula asserting a falsity about an x is made to assert a falsity about the Gödel set associated with that very formula, which leads to a contradiction. Now, in the case of the definition of truth given above, the method of diagonalization cannot be used to derive paradoxical results. The reason is, indeed, that the relations of satisfaction and truth in V_{θ_1} have been defined only for formulas relativized to V_{θ_1}. By contrast, the definitions themselves require and explicitly mention the existence of V_{θ_1}. It follows that the defined relations hold only for formulas or sentences that do not involve quantification over a domain that contains V_{θ_1} as a member. This implies that no formula in the list of formulas to which these relations are applicable may itself involve the relational expressions '$Sat_s(u, V_{\theta_1}, b)$' or '$Tr_s(u, V_{\theta_1})$'. Therefore, the diagonalization fails, and no contradiction ensues.

Does the ZFC+A-Ina relation $Tr_s(\ulcorner \varphi^{V_{\theta_1}} \urcorner, V_{\theta_1})$ pass the test of adequacy imposed by the convention T (C-T)? As we know, the convention T requires of the truth definition that it should be able to *prove* all the biconditionals of the form: $Tr(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$, where φ is a sentence of the language for which the truth is defined, i.e., φ itself does not contain the truth predicate. Adjusted to the shape of our truth definition (D-Tr$_s$), the convention T becomes the requirement:

$$\text{ZFC+A-Ina} \vdash Tr_s(\ulcorner \varphi^{V_{\theta_1}} \urcorner, V_{\theta_1}) \leftrightarrow \varphi^{V_{\theta_1}}.$$

So is ZFC+A-Ina capable of proving any such biconditional?

To convince ourselves that our definition meets the requirement, we will merely provide an example, not a full proof. Consider once again the familiar sample sentence: $\exists v_1 \exists v_2 (v_1 \in v_2)$. Relativized to V_{θ_1} it becomes: $\exists x \in V_{\theta_1}$ $\exists y \in V_{\theta_1} (x \in y)$. To maximize the simplicity of the argument, we will make our life easier and assume that this sentence was formed in only three steps along the sequence: $v_1 \in v_2$, $\exists v_2 \in V_{\theta_1} (v_1 \in v_2)$, and $\exists v_1 \in V_{\theta_1} \exists v_2 \in V_{\theta_1} (v_1 \in v_2)$. Owing to the presence of 'V_{θ_1}', the Gödel set for these formulas will be rather complicated in structure, therefore, we will use merely the corner quotes to denote it. Let b be any assignment $b : \rho(u) \mapsto V_{\theta_1}$, where u is the

Gödel number of the appropriate formula. What we aim to derive is the biconditional:

$$\forall b\big(Sat_s(\ulcorner \exists x \in V_{\theta_1} \exists y \in V_{\theta_1}(x \in y)\urcorner, V_{\theta_1}, b)\big) \leftrightarrow \exists x \in V_{\theta_1} \exists y \in V_{\theta_1}(x \in y).$$

The easiest way of showing that the biconditional holds in ZFC+A-Ina is to start with the atomic formula, and then continue by prefixing quantifiers. When does $Sat_s(\ulcorner v_1 \in v_2\urcorner, V_{\theta_1}, b)$ hold? If and only if $b(1) \in b(2)$. And when does it hold for all assignments b? The reply is, indeed, in the biconditional: $\forall b(Sat_s(\ulcorner v_1 \in v_2\urcorner, V_{\theta_1}, b)) \leftrightarrow \forall b(b(1) \in b(2))$. Let us move on to the second step and prefix the quantifier: $Sat_s(\ulcorner \exists v_2 \in V_{\theta_1}(v_1 \in v_2)\urcorner, V_{\theta_1}, b)$. This holds if and only if $\exists y \in V_{\theta_1}(b(1) \in y)$, which gets us to the following biconditional: $\forall b(Sat_s(\ulcorner \exists v_2 \in V_{\theta_1}(v_1 \in v_2)\urcorner, V_{\theta_1}, b)) \leftrightarrow \forall b \exists y \in V_{\theta_1}(b(1) \in y)$. If we repeat the same procedure in the remaining third formation step of our sample sentence, we arrive at the final biconditional: $\forall b(Sat_s(\ulcorner \exists v_1 \in V_{\theta_1} \exists v_2 \in V_{\theta_1}$ $(v_1 \in v_2)\urcorner, V_{\theta_1}, b)) \leftrightarrow \forall b \exists x \in V_{\theta_1} \exists y \in V_{\theta_1}(x \in y)$. But now the quantifier '$\forall b$' on the right-hand side of the equivalence arrow has become vacuous as there is no occurrence of 'b' within its scope, so it may be dropped. Thus we end up with the biconditional we wished to prove. As any sentence of $\mathcal{L}_{\mathrm{ZFC}}$ has been formed in a finite number of steps, it will always be in principle possible to apply a procedure similar to the one we have just suggested.

So we are free to consider ZFC+A-Ina to be an adequate theory of truth for the $\mathcal{L}_{\mathrm{ZFC}}$ sentences $\varphi^{V_{\theta_1}}$. Yet, we might want from our relation $Tr_s(\ulcorner \varphi^{V_{\theta_1}} \urcorner, V_{\theta_1})$ a little bit more than just to meet the adequacy requirement imposed by the convention T. To understand what, let us divide the sentences of $\mathcal{L}_{\mathrm{ZFC}}$ into three broad categories. (Naturally, this division presupposes that ZFC is consistent; otherwise the three categories collapse into a single one of theorems.) In the first group are the theorems of ZFC and their negations; to the second group, a polar opposite to the first one, belong the sentences about whose truth or falsity ZFC does not provide any evidence; finally, in the third group, there are the sentences that are not theorems of ZFC but can be, under certain assumptions, shown to be true and their negations false. What is the relationship between the sentences that are Tr_s in V_{θ_1} and these three broad categories of sentences? Specifically, does the property of being Tr_s in V_{θ_1} hold of the sentences belonging to both the first and the second category?

Let us first have a quick look at the second category. These are sentences that are independent of ZFC and whose truth or falsity cannot in any way be established on the grounds of the axioms of ZFC. Examples of such independent statements are the continuum hypothesis or the claims concerning the existence of large ordinals and cardinals. Obviously, our relation $Tr_s(\ulcorner \varphi^{V_{\theta_1}} \urcorner, V_{\theta_1})$ is not of any help with regard to establishing the truth value of such statements. It cannot be of any use in showing the truth or falsity of statements including large sets because such statements would

get translated into relativizations whose truth values would be determined by factors other than those pertaining to the existence or non-existence of the large sets. On the other hand, the truth or falsity of statements such as the continuum hypothesis is, in this respect, a totally different matter as it concerns sets deep within the lower segments of the cumulative hierarchy, in any case well below V_{θ_1}. Therefore, the relation $Tr_s(\ulcorner \varphi^{V_{\theta_1}} \urcorner, V_{\theta_1})$ should in principle apply to sentences expressing the continuum hypothesis (CH) or their negations. If these sentences are deemed meaningful and the law of the excluded middle holds, we should have either $Tr_s(\ulcorner \mathrm{CH}^{V_{\theta_1}} \urcorner, V_{\theta_1})$ or $Tr_s(\ulcorner \neg \mathrm{CH}^{V_{\theta_1}} \urcorner, V_{\theta_1})$. Fortunately, we do not have to immerse in the debate surrounding the continuum hypothesis.[22] What is at stake at the moment is merely the sufficiency of our definition of truth. In general, we do not require that the definition of truth should be able to help us establish the truth or falsity in V_{θ_1} of every single sentence $\varphi^{V_{\theta_1}}$ of $\mathcal{L}_{\mathrm{ZFC}}$; the purpose of the definition of truth is not to perform an exhaustive assortment of all sentences $\varphi^{V_{\theta_1}}$ into the sets True and False. Unless the theories we are dealing with, i.e., ZFC and ZFC+A-Ina, provide some grounds for a decision in the given matter, there is no reason why a theory of truth based on them should do any better.

There is, nevertheless, one more important aspect to this problem that needs to be pointed out. It has to do with the difference between the first-order and second-order theories, i.e., between ZFC and ZFC_2. As the logic of ZFC is complete, whatever is provable in ZFC is also a logical consequence of ZFC. ZFC is neither categorical, nor quasi-categorical (recall our discussion of the Löwenheim-Skolem theorem in section 3.4), i.e., it has non-isomorphic models even of the sentences $\varphi^{V_{\theta_1}}$. It follows that the continuum hypothesis, being independent of ZFC and not being a logical consequence of ZFC, will be true in some models of ZFC and false in others. On the other hand, the situation of ZFC_2 is very much different. Its logic is incomplete, so ZFC_2 has logical consequences that are not provable in it. At the same time, it is quasi-categorical, i.e., provided that there are no urelements, all its models are isomorphic up to a given level V_α. This means that they all agree on the lower levels. In particular, they have to agree on the dilemma of the continuum hypothesis: they all have to decide it in one way or the other. The trouble is, of course, that we do not know in which way the models of ZFC_2 solve this problem. As van Benthem and Doets [1983] put it, ZFC_2

[22] The question of the truth value of the continuum hypothesis is, indeed, a very intricate one, and it can be argued that although the proofs by Gödel and Cohen that both the continuum hypothesis and its negation are consistent with ZFC answered the problem figuring as number one (1a) in Hilbert's famous list (cf. Hilbert [1900]), the whole matter still sticks out as an unresolved issue cutting deep dividing lines between various currents in the philosophy of mathematics. Roughly, the attitudes range from considering the continuum hypothesis to be true (e.g., Cantor, Hilbert [1926], pp. 384–392), false (e.g., Gödel [1947], Cohen [1966], pp. 150–152) or not sufficiently determined to be either (e.g., Skolem [1923], p. 299n., who points to the fact that ZFC, being first-order, is not categorical).

'knows the answers—unfortunately, we're not able to figure out exactly what it knows' (p. 296).

Let us return to our three categories of sentences, and let us have a look at the sentences belonging to the first category, i.e., the theorems and their negations, which can be shown—unless ZFC is inconsistent—to be true or false, respectively. That is, theorems are demonstrably true, therefore we want them to be Tr_s in V_{θ_1}. (Note that we no longer need to relativize φ because we already know that any theorem of ZFC is true in the model with the domain V_{θ_1}, hence for any theorem φ of ZFC, $\varphi \leftrightarrow \varphi^{V_{\theta_1}}$. This makes the relativization superfluous.) Let φ be a theorem of ZFC, and let us ask: does the relation $Tr_s(\ulcorner\varphi\urcorner, V_{\theta_1})$ coincide with the informal relation of φ's being true in V_{θ_1}? This question can be immediately answered in the positive. As ZFC+A-Ina is an extension of ZFC, anything provable in ZFC will also be provable in ZFC+A-Ina. In addition, we take ZFC+A-Ina to be able to prove all the biconditionals required by the convention T. From this it follows that once ZFC+A-Ina can prove a ZFC theorem φ, it is also able to prove the required biconditional $Tr_s(\ulcorner\varphi\urcorner, V_{\theta_1}) \leftrightarrow \varphi$. Hence $Tr_s(\ulcorner\varphi\urcorner, V_{\theta_1})$ holds. So we can conclude that the following obtains: $\forall u(Prov(u) \rightarrow Tr_s(u, V_{\theta_1}))$, where '$u$' is the Gödel number of a sentence of $\mathcal{L}_{\mathrm{ZFC}}$ and '$Prov(u)$' expresses the provability property for ZFC.[23] We will abbreviate this implication as '$Tr_s(\mathrm{ZFC}, V_{\theta_1})$'.

Finally, there are sentences belonging to the third category. These are independent of ZFC but still can be shown to be true, provided that we stick to the normal interpretation of '$\in$' as the membership relation (and that we interpret the sentences of $\mathcal{L}_{\mathrm{ZFC}}$ as true and false at all[24]). It follows from Gödel's incompleteness theorems that no consistent recursively axiomatizable theory that is capable of capturing (case-by-case proving) all primitive recursive functions as functions[25] is complete. ZFC is such a theory, hence

[23]The provability property $Prov(u)$ would be defined as $\exists x(Prf(x, u))$ where $Prf(x, u)$ holds if x is the Gödel set of a ZFC proof of the Gödel set u associated with a sentence of $\mathcal{L}_{\mathrm{ZFC}}$. However, we have not introduced any technique of arithmetization to code proofs, i.e., sequences of formulas. Since this can be done, and there is nothing problematic about it, we will merely assume that we have the provability property at our disposal to save us room and effort.

[24]For a discussion of the relationship between Gödel's incompleteness theorems and mathematical truth see Franzén [2005], pp. 28–33.

[25]The terminology does not seem to be fixed here. I follow Smith [2007], pp. 99–105. A one-place function f is CAPTURED (case-by-case proved) by a formula $\varphi(x, y)$ in theory T if and only if for any m, n:

$$\text{if} f(m) = n, \text{then } \mathrm{T} \vdash \varphi(\overline{m}, \overline{n}),$$
$$\text{if} f(m) \neq n, \text{then } \mathrm{T} \vdash \neg\varphi(\overline{m}, \overline{n}).$$

If, moreover:

$$\text{for every } m, \mathrm{T} \vdash \exists! y \varphi(\overline{m}, y),$$

where '$\exists!$' is the uniqueness quantifier, the function f is CAPTURED AS A FUNCTION. The concept of capturing a function can be extended to properties and relations. We say that

it is incomplete. It is important to realize that this type of incompleteness is rather different from the one connected with the questions of the continuum hypothesis or the existence of large sets. Whereas those questions may be, at least in theory, simply pushed away by adding new axioms that decide them by fiat, the incompleteness proved by Gödel is more fundamental in the sense that it "sticks". It is unremovable; no matter how many new axioms we add, provided that we do not give up the consistency or the recursive axiomatizability, the theory will remain incomplete. However, as it has been said, the sentences serving as witnesses of the irreparable incompleteness of ZFC are characterized by the remarkable attribute that, under the aforementioned conditions, they are true and can, under certain rather basic conditions, also be shown to be true. The question then arises: does our relation $Tr_s(\ulcorner\varphi^{V_{\theta_1}}\urcorner, V_{\theta_1})$ apply to them? Can they be shown to be Tr_s in V_{θ_1}?

In what follows, we will demonstrate that ZFC+A-Ina is able to prove both the Gödel sentence exemplifying the incompleteness of ZFC (which is the subject matter of Gödel's first incompleteness theorem) and the sentence asserting the consistency of ZFC (exemplifying Gödel's second incompleteness theorem). We will begin with the canonical Gödel sentence. Let '$Prov(u)$' be the provability predicate for ZFC; then the Gödel sentence, G, says: $\neg Prov(\ulcorner\neg Prov(x)\urcorner)$. Now the following holds: ZFC $\vdash G \leftrightarrow \neg Prov(\ulcorner G\urcorner)$.[26] To prove the Gödel sentence for ZFC in ZFC+A-Ina, we only need to recall (Fact 1) that ZFC+A-Ina can prove that $Tr_s(\text{ZFC}, V_{\theta_1})$, i.e., $\forall u(Prov(u) \rightarrow Tr_s(u, V_{\theta_1}))$, and (Fact 2) that it is capable of proving all the biconditionals of the convention T. Then we proceed as follows. First, by Fact 1 ZFC+A-Ina $\vdash Prov(\ulcorner G\urcorner) \rightarrow Tr_s(\ulcorner G\urcorner, V_{\theta_1})$. Secondly, by Fact 2 ZFC+A-Ina $\vdash Prov(\ulcorner G\urcorner) \rightarrow G$. However, G is equivalent to $\neg Prov(\ulcorner G\urcorner)$; as $Prov(\ulcorner G\urcorner)$ is the negation of G, we can substitute, and we obtain, thirdly, $\neg G \rightarrow G$. This implication is true only if G is true. Hence ZFC+A-Ina $\vdash G$.[27]

Let us move on to the example of a sentence asserting the consistency of a theory. A theory is consistent if there is no contradiction among its theorems.

---

a formula $\varphi(x_1, \ldots, x_n)$ captures a relation R in T, if for any $m_1, \ldots, m_n$:

$$\text{if } R(m_1, \ldots, m_n), \text{then T} \vdash \varphi(\overline{m}_1, \ldots, \overline{m}_n),$$
$$\text{if } \neg R(m_1, \ldots, m_n), \text{then T} \vdash \neg\varphi(\overline{m}_1, \ldots, \overline{m}_n).$$

Similarly to the concept of definability (defined on p. 89), the purpose of the concept of capturability is to represent functions or properties and relations of natural numbers by means of formulas. Yet, whereas in the case of definability we deal with the *truth* of the defining formula in a model, here we consider the *provability* of the capturing formula in a theory.

[26] This statement follows from the fixed point theorem, or the diagonalization lemma. The lemma is discussed in section 5.2, p. 120.

[27] This proof is based upon the proof for PA in Ketland [1999], pp. 86–88. The proof that follows is also to be attributed to Ketland, op. cit., pp. 81–82.

Take the sentence '$0 = 1$' as a sample contradiction. Then ZFC is consistent, if $\neg Prov(\ulcorner 0 = 1 \urcorner)$. The second incompleteness theorem, as applied to ZFC says that unless ZFC is inconsistent, ZFC $\nvdash \neg Prov(\ulcorner 0 = 1 \urcorner)$. Now in ZFC+A-Ina it can easily be proved that $\neg Prov(\ulcorner 0 = 1 \urcorner)$ as follows. First, by Fact 1 we have that ZFC+A-Ina $\vdash Prov(\ulcorner 0 = 1 \urcorner) \rightarrow Tr_s(\ulcorner 0 = 1 \urcorner, V_{\theta_1})$. Secondly, by Fact 2 we obtain that ZFC+A-Ina $\vdash Prov(\ulcorner 0 = 1 \urcorner) \rightarrow 0 = 1$. But of course, ZFC+A-Ina can prove that $0 \neq 1$. Hence ZFC+A-Ina $\vdash \neg Prov(\ulcorner 0 = 1 \urcorner)$. This, of course, does not mean anything else than that ZFC+A-Ina proves the consistency of ZFC. To sum up, the statements exemplifying the incompleteness and consistency of ZFC can be proved, with the help of the relation $Tr_s(\ulcorner \varphi^{V_{\theta_1}} \urcorner, V_{\theta_1})$, in ZFC+A-Ina. Therefore, we may conclude that the relation $Tr_s(\ulcorner \varphi^{V_{\theta_1}} \urcorner, V_{\theta_1})$ performs as desired with respect to the sentences of the third category.

We have seen that to define the concept of truth for (the Gödel sets of) the sentences of $\mathcal{L}_{\mathrm{ZFC}}$ relativized to V_{θ_1} it suffices to add to ZFC an additional axiom asserting the existence of the level indexed by an inaccessible ordinal, e.g., V_{θ_1}. Besides its being able to prove that there exists a model of ZFC, the resulting theory—in our case ZFC+A-Ina—is not conservative over ZFC even with respect to the relativized sentences: it is capable of proving the Gödel sentence for ZFC as well as the sentence expressing the consistency of ZFC, which are both unprovable in ZFC.

4.6 Truth: a Class Form

If we want to define the relations of truth and satisfaction for $\mathcal{L}_{\mathrm{ZFC}}$, there is yet another path we could choose to follow. Instead of adding new axioms to ZFC and relativizing the sentences of $\mathcal{L}_{\mathrm{ZFC}}$, we could turn to the second-order set theory such as ZFC_2, and exploit the availability of second-order entities, namely classes. In the present section, we will examine this alternative.

$\mathcal{L}_{\mathrm{ZFC}_2}$, the language of ZFC_2, of course, differs from $\mathcal{L}_{\mathrm{ZFC}}$, which we were dealing with in the preceding sections. As we know, $\mathcal{L}_{\mathrm{ZFC}_2}$ is an extension of $\mathcal{L}_{\mathrm{ZFC}}$, i.e., it is able to express everything $\mathcal{L}_{\mathrm{ZFC}}$ is able to plus something more. In section 4.1, we introduced the technique of arithmetization of syntax only for $\mathcal{L}_{\mathrm{ZFC}}$, saying that we would not bother with providing details of the extension of it also for $\mathcal{L}_{\mathrm{ZFC}_2}$. As it was defined only for formulas of $\mathcal{L}_{\mathrm{ZFC}}$, it can be employed here without any change. If we wanted to include formulas containing second-order variables, we would have to, of course, extend the technique of arithmetization so that it covered the whole of $\mathcal{L}_{\mathrm{ZFC}_2}$. As we have already said, this can be done without a difficulty but, if needed, we will simply assume that such a system is available.

It turns out that, for the definitions of the class variants Sat_c and Tr_c of satisfaction and truth, our old faulty definitions (D-S*), (D-Sat*) and

(D-Tr*) may be taken over with only minimal changes, so our progress will be quick. First of all, the definition (D-Fm) on p. 83 will be adopted without any change. It just needs to be born in mind that the formation function s, which is one of the three terms of Fm, no longer describes the formation (in $n+1$ steps) of the Gödel set of any $\mathcal{L}_{\mathrm{ZFC}_2}$ formula but only of any $\mathcal{L}_{\mathrm{ZFC}}$ formula. The individual clauses of the disjunction part of (D-Fm) rule out all formulas containing the class variables. The relation $Fm(u,s,n)$ thus holds only between the natural number n, the formation function s and the Gödel set of any formula belonging to the $\mathcal{L}_{\mathrm{ZFC}}$ fragment of $\mathcal{L}_{\mathrm{ZFC}_2}$.

The definition (D-S*) can be taken without any apparent change at all. Despite its remaining the same, let us make it readily available, and restate once again at least its last clause:

$$\exists i < \omega \exists l < k[s(k) = \langle 4, i, s(l)\rangle \wedge t(k) = \{a \mid \exists x(a(i/x) \in t(l))\}]. \qquad \text{(D-S}_{\mathrm{c}}\text{)}$$

(Take (D-S$_{\mathrm{c}}$) to designate the whole definition, not just this single clause.) Although not requiring any visible intervention, the definition (D-S$_{\mathrm{c}}$) has to be reinterpreted. The value that t assigns to k must be a class. We already saw that it cannot be a set in section 4.2 but at that time we did not have any other option. Now we do have one: the abstraction operator $\{\mid\}$ has to be read as 'the class of a such that etc.' This is to say that the range of the function t is a class of classes, which makes it no longer possible to regard it as a set. That is, t itself needs to be recognized as a class.

A note on the notation. We will follow the habit of using the lower-case letters for function variables to make them clearly distinguishable from relation variables, which involve capital letters. Thus our notation does not show whether a function variable or a relation variable is to be read as ranging over sets or classes. To avoid ambiguity, we will always clearly state if a variable is to be read as ranging over classes. Unless explicitly stated, every function and relation variable is to be read as ranging over sets.

Now, with the relation S_c in hand, the definitions of satisfaction and truth are straightforward. The definition of the class version of satisfaction almost exactly corresponds to the faulty (D-Sat*):

$$\begin{aligned} Sat_c(u,b) \leftrightarrow_{Def.}\ & \\ & \exists t \exists s, n \in V_\omega \big(Fm(u,s,n) \wedge Func(t) \wedge \qquad \text{(D-Sat}_{\mathrm{c}}\text{)} \\ & dom(t) = n+1 \ \wedge\ b \in t(n) \ \wedge\ \forall k \leq n(S_c(k,t,s))\big), \end{aligned}$$

where t ranges over classes. Eventually, the definition of the class version of the property of being true is also almost identical to (D-Tr*). Still, let us write it down, for the sake of completeness:

$$Tr_c(u) \leftrightarrow_{Def.} Sent(u) \wedge \forall b(Sat_c(u,b)). \qquad \text{(D-Tr}_{\mathrm{c}}\text{)}$$

Here b, an individual assignment, is to be interpreted as a set. It is, so to speak, the structural characteristic determining which of the sets are assign-

ments that can only be interpreted in terms of classes but not individual assignments. Having laid down (D-Tr$_c$), we are done.

How does the property Tr_c fare with respect to the convention T? The direction $\mathrm{ZFC}_2 \vdash Tr_c(\ulcorner\varphi\urcorner) \rightarrow \varphi$ is unproblematic. We know that the property Tr_c can hold only of the sentences of $\mathcal{L}_{\mathrm{ZFC}_2}$ that are also sentences of $\mathcal{L}_{\mathrm{ZFC}}$. To show that this implication is true, we can use the same method as in section 4.5. On the other hand, the other direction in its full generality does not hold. It is simply not the case that $\mathrm{ZFC}_2 \vdash \varphi \rightarrow Tr_c(\ulcorner\varphi\urcorner)$. The reason is obvious: φ is any sentence of $\mathcal{L}_{\mathrm{ZFC}_2}$, i.e., it can contain second-order variables, for which the property Tr_c is not defined. So we have to restrict the scope of sentences that are required to pass the test of the convention T to those sentences of $\mathcal{L}_{\mathrm{ZFC}_2}$ that are at the same time sentences of $\mathcal{L}_{\mathrm{ZFC}}$. Note that this is not a drawback that would disqualify our definition (D-Tr$_c$) as what we have been after since the beginning was only a partial truth definition precisely for the sentences of $\mathcal{L}_{\mathrm{ZFC}}$. It is merely a technical obstacle that needs to be resolved. Anyway, we have, basically, two options. Either we can impose an external condition on the convention T restricting its applicability only to $\mathcal{L}_{\mathrm{ZFC}}$ sentences. Or we can come up with a "structural" property definable in ZFC_2 that will make it possible to insert an additional condition into the convention T. As we wish to show that a partial concept of truth is definable within ZFC_2, we need to opt for the latter. Actually, to find such a "structural" property is nothing difficult, and a number of approaches might be taken. The general method would be to use or define a property holding of just those Gödel sets that are associated with sentences that do not quantify over higher-order entities. We have used in (D-Tr$_c$) the property *Sent* (which we have not defined explicitly but only assumed that it can be defined in a manner similar to that of *Form* on p. 83). This property *Sent* has been (assumed to be) defined already within ZFC, so it is not a general property holding of the Gödel set of any sentence of $\mathcal{L}_{\mathrm{ZFC}_2}$, but only of a sentence of $\mathcal{L}_{\mathrm{ZFC}}$. To make absolutely clear its scope, let us designate it as '$Sent_{\mathrm{ZFC}}$'. Employing this property, the convention T may be transformed into the requirement that the following schema of $\mathcal{L}_{\mathrm{ZFC}_2}$ should hold for every sentence φ of $\mathcal{L}_{\mathrm{ZFC}_2}$:

$$\mathrm{ZFC}_2 \vdash Sent_{\mathrm{ZFC}}(\ulcorner\varphi\urcorner) \rightarrow (Tr_c(\ulcorner\varphi\urcorner) \leftrightarrow \varphi). \qquad \text{(C-Tr}_c\text{)}$$

Showing that ZFC_2 meets the condition imposed by C-Tr$_c$ does not essentially differ from what we did in section 4.5. So we can conclude that Tr_c satisfies the requirement on a partial concept of truth.

The other arguments of section 4.5 also need not be repeated. Let us just state as plain facts that, again, the method of diagonalization cannot be used to derive an immediate contradiction, and that ZFC_2 is strong enough to prove that the theorems of ZFC are Tr_c, as well as the consistency of the first-order system of ZFC.

Let us add, as an interesting aside, that the relations of satisfaction and truth for $\mathcal{L}_{\mathrm{ZFC}}$ can be defined also in NBG and MK set theories. However, it was demonstrated by Mostowski [1950] that NBG (unless it is inconsistent) cannot—in spite of its expressive resources that enable it to formulate the definition of truth in such a way that the convention T is satisfied—prove the general statement that every theorem of ZFC is true, and consequently, that ZFC is consistent. The reason for this is obvious. As we know, NBG is a conservative extension of ZFC, i.e., it cannot prove more facts concerned purely with sets than ZFC is able to. The statement that ZFC is consistent can be formulated as a statement purely about sets. Therefore, NBG cannot prove that ZFC is consistent unless it is itself (and ZFC) inconsistent. Thus NBG is only a very weak theory of truth for $\mathcal{L}_{\mathrm{ZFC}}$. In MK, on the other hand, truth for $\mathcal{L}_{\mathrm{ZFC}}$ is definable.

4.7 Truth for Sentences of Restricted Complexity

In sections 4.4, 4.5 and 4.6, we aimed at developing the strategy based on the idea of restricting the domain of objects to which the relations of satisfaction and truth are applicable. Now we will examine the other strategy mentioned at the outset of section 4.4, which purports to define the partial truth for sentences whose complexity has been limited. In carrying out this task, the Lévy hierarchy of formulas will become a vital tool. Once again, we will focus our attention primarily on the first-order system of ZFC.

First of all, we need to expand the coding scheme introduced in section 4.1 by adding the following clause for the bounded existential quantifier:

$$\ulcorner \exists v_i \in v_j \varphi \urcorner \text{ is } \langle 5, i, j, \ulcorner \varphi \urcorner \rangle.$$

(Strictly speaking, this clause is not necessary. However, without it we would have to define—in order to distinguish between bounded and unbounded quantification—a complex syntactic property that would hold of all and only those formulas that contain bounded quantifiers. It is, of course, rather more expedient simply to expand the coding scheme, and avoid going through a lot of tedious labour needed to achieve this.)

Now we will define the auxiliary relation $Satfun(f, n)$ which is to hold if and only if f is a two-place satisfaction function $f(u, b)$, where u is the Gödel set of a formula φ belonging to $\Sigma_n^{\mathrm{ZFC}} \cup \Pi_n^{\mathrm{ZFC}}$ and b is an assignment of values to the free variables occurring in φ. The range of values of f will be $\{0, 1\}$: $f(u, b) = 1$ if b satisfies u and $f(u, b) = 0$ if it does not. The definition of $Satfun$ spelled out in full looks rather complex:[28]

[28] The definition (D-Sfu) is modeled on the definition first given in Lévy [1965], p. 22. Cf. also Drake [1974], p. 98.

$$
\begin{aligned}
Satfun(f,n) \leftrightarrow_{Def.}\ & Func(f) \wedge rng(f) = \{0,1\} \wedge \\
& \forall \langle u,b\rangle \in dom(f)\big(degree(u) \leq n \wedge dom(b) = freev(u) \wedge \\
& \big[\exists i,j < \omega\big([u = \langle 0,i,j\rangle \wedge (f(u,b) = 1 \leftrightarrow b(i) = b(j))] \vee \\
& \quad [u = \langle 1,i,j\rangle \wedge (f(u,b) = 1 \leftrightarrow b(i) \in b(j))]\big) \vee \\
& \quad \exists v,w\big([u = \langle 2,v,w\rangle \wedge (\langle v,b\rangle \in dom(f) \wedge \langle w,b\rangle \in dom(f) \wedge \\
& \qquad (f(u,b) = 1 \leftrightarrow (f(v,b) = 1 \vee f(w,b) = 1)))] \vee \\
& \quad [u = \langle 3,v\rangle \wedge \langle v,b\rangle \in dom(f) \wedge (f(u,b) = 1 \leftrightarrow f(v,b) = 0)]\big) \vee \\
& \exists i,j < \omega \exists v[u = \langle 5,i,j,v\rangle \wedge \forall x \in b(j)(\langle v, b(i/x)\rangle \in dom(f)) \wedge \\
& \qquad (f(u,b) = 1 \leftrightarrow \exists x \in b(j)(f(v,b(i/x)) = 1))] \vee \\
& \exists i < \omega \exists v[u = \langle 4,i,v\rangle \wedge \forall j < \omega(u \neq \langle 5,i,j,v\rangle) \wedge \\
& \qquad (f(u,b) = 1 \leftrightarrow \exists x(Sat_d(v, n-1, b(i/x))))]\big).
\end{aligned}
\tag{D-Sfu}
$$

If $n = 0$, the last clause (i.e., the clause for $u = \langle 4,i,v\rangle$) is omitted. The expressions '$rng(f)$', '$degree(u)$' and '$freev(u)$' mean 'the range of f', 'the degree of the formula associated with u' and 'the set of free variables occurring in the formula associated with u', respectively. These functions are all definable in ZFC. Note that the assignment b is a function from the set of free variables occurring in a formula into the universe of sets, i.e., its range is left unrestricted.

The satisfaction relation $Sat_d(u,n,b)$ is defined as follows:

$$
Sat_d(u,n,b) \leftrightarrow_{Def.} \exists f(Satfun(f,n) \wedge \langle u,b\rangle \in dom(f) \wedge f(u,b) = 1). \tag{D-Sat$_\mathrm{d}$}
$$

In plain words, $Sat_d(u,n,b)$ holds if and only if there is a satisfaction function f satisfying the Gödel set u associated with a formula belonging to $\Sigma_n^{\mathrm{ZFC}} \cup \Pi_n^{\mathrm{ZFC}}$ under the assignment b. (We use the lower index '$_d$' to indicate that Sat_d is the relation of satisfaction defined only for formulas of a specific degree.) The relation $Sat_d(u,n,b)$ is $\Delta_{n+1}^{\mathrm{ZFC}}$.[29] Employing the relation of satisfaction, the property of truth can be defined simply as:

$$
Tr_d(u,n) \leftrightarrow_{Def.} Sent(u) \wedge \forall b(Sat_d(u,n,b)). \tag{D-Tr$_\mathrm{d}$}
$$

Roughly speaking, (D-Tr$_\mathrm{d}$) says that the Gödel set u associated with a $\mathcal{L}_{\mathrm{ZFC}}$ sentence that belongs to $\Sigma_n^{\mathrm{ZFC}} \cup \Pi_n^{\mathrm{ZFC}}$ is Tr_d if and only if it is satisfied by all assignments. It is easy to see that the property Tr_d is also $\Delta_{n+1}^{\mathrm{ZFC}}$ (notice that b, by (D-Sat$_\mathrm{d}$), is bounded). Sometimes we will use again the fancy notation $\models_d^n \ulcorner\varphi\urcorner\,[b]$ instead of $Sat_d(\ulcorner\varphi\urcorner, n, b)$, and $\models_d^n \ulcorner\varphi\urcorner$ instead of $Tr_d(\ulcorner\varphi\urcorner, n)$.

Having defined the partial relations of satisfaction and truth, we need to ask: What is it that makes the definitions acceptable? After all, we have

[29] The alternative definition of $Sat_d(u,n,b)$ can be given by the equivalent formula: $u \in \Sigma_n^{\mathrm{ZFC}} \cup \Pi_n^{\mathrm{ZFC}} \wedge$ 'b is an assignment' $\wedge \forall f(Satfun(f,n) \wedge \langle u,b\rangle \in dom(f) \rightarrow f(u,b) = 1)$. Cf. Drake [1974], p. 98. The full proof is in Lévy [1965], pp. 24–25.

pointed out that the range of the assignment function b is unrestricted. How does it come about that the relation $Satfun$, which is is the collection of all the ordered pairs $\langle\langle\langle u, b\rangle, 1 \text{ or } 0\rangle, n\rangle$, i.e., which contains all the functions f, can still be interpreted as a set? To explain this, we need to employ the notion of absoluteness defined in section 4.3. Recall that the universe of sets we are considering, namely V, is the universe of transitive sets. Suppose that φ is a Δ_0^{ZFC} formula that is satisfiable in a transitive subdomain M of V. Then, as we know, φ is absolute for M. Why is this fact significant? Take, as an elementary example, an atomic Δ_0^{ZFC} formula $v_1 \in v_2$. This formula is satisfied by any assignment that assigns to the variables objects a and b such that $a \in b$. In the extreme case, it is satisfiable in a domain containing just two sets such that one is a member of the other. In general, Δ_0^{ZFC} formulas are satisfiable in relatively restricted subdomains of V, and there is nothing in these formulas that forces us to go beyond such subdomains. Absoluteness guarantees that formulas that are satisfied in M will not cease to be satisfied in V. Still more importantly, this also works in the other direction. Once a Δ_0^{ZFC} formula is satisfied in V by some sets, we may take some or all of these sets as a domain of satisfaction of the formula. The key point to realize is that formulas with only free variables and, for obvious reasons, formulas with only bounded quantifiers do not force us to run through the entire universe V when finding out whether they are satisfied by an assignment or not. Therefore, when dealing only with Δ_0^{ZFC} formulas, we may consider only a subdomain of the universe, i.e., a set. However, we need to start adding unbounded quantifiers. This complicates the picture just drawn since the very purpose of the unbounded quantifier is to range over every single object in V. But even here we can interpret the collection of all the assignments as a set. Just recall what goes on when a variable gets bound, e.g., in the Δ_0^{ZFC} formula $v_1 \in v_2$. Adding the existential quantifier, we obtain the Σ_1^{ZFC} formula $\exists v_2(v_1 \in v_2)$. The assignments b that assigned objects to the free variables v_1 and v_2 will be replaced by assignments that assign objects to the remaining free variable v_1 but at the place of the newly bound variable they will just contain x. Symbolically, $\models_d^1 \ulcorner \exists v_2(v_1 \in v_2) \urcorner [b]$ if and only if $\exists x(\models_d^0 \ulcorner v_1 \in x \urcorner [b])$.[30] It is thus not required that the assignments should exhaust the entire universe.

It is obvious that the definition (D-Tr_d) escapes the contradiction described in section 4.2. As the truth predicate $Tr_d(\ulcorner \varphi \urcorner, n)$, which itself is $\Delta_{n+1}^{\text{ZFC}}$, is always defined only for formulas $\Sigma_{m \leq n}^{\text{ZFC}} \cup \Pi_{m \leq n}^{\text{ZFC}}$, the diagonalization fails, and no formula can be made to assert truth about the Gödel set associated with its own negation. Now the basic idea the definition (D-Tr_d) is founded upon is more or less identical to that underlying (D-Tr_s); it merely replaces the restriction imposed on the domain with the restriction imposed on the complexity of formulas. Therefore, it is unnecessary to go through

[30] For a generalization of this idea, cf. Kanamori [2003], p. 6.

the arguments supporting the claim that ZFC satisfies the convention T, i.e., that for every $\Sigma^{\mathrm{ZFC}}_{m\leq n} \cup \Pi^{\mathrm{ZFC}}_{m\leq n}$ formula φ ZFC $\vdash Tr_d(\ulcorner\varphi\urcorner, n) \leftrightarrow \varphi$. We will also not show but merely state that, making use of the property $Tr_d(\ulcorner\varphi\urcorner, n)$, ZFC is able to prove that every theorem of ZFC of degree $\leq n$ is true as well as the consistency of the set of all the formulas belonging to $\Sigma^{\mathrm{ZFC}}_{m\leq n} \cup \Pi^{\mathrm{ZFC}}_{m\leq n}$, i.e., ZFC can prove the consistency of ZFC restricted to the sentences of degree $\leq n$.[31] Because there is no upper bound on the degree of theorems of ZFC, there is no way of restricting the formulas or sentences of $\mathcal{L}_{\mathrm{ZFC}}$ so that we could circumvent them, and prove that they are all Tr_d for some n, and that ZFC is consistent.

This feature of ZFC can be described in terms of ω-incompleteness. The formula $Tr_d(\ulcorner\varphi\urcorner, n)$ can be used to define, for each degree n, a set of true sentences φ of degree $\leq n$. This can achieved, for instance, by means of the formula $t_n = \{\ulcorner\varphi\urcorner \mid Tr_d(\ulcorner\varphi\urcorner, \overline{n})\}$. In this way we obtain the sequence of sets $t_0, t_1, \ldots, t_n, \ldots$, each a set of partial truth. It can be shown that ZFC, unless it is inconsistent, can prove that there is such a truth set t_n for each particular n, i.e., ZFC $\vdash \exists x(x = \{\ulcorner\varphi\urcorner \mid Tr_d(\ulcorner\varphi\urcorner, \overline{n})\})$ for any n, but that it cannot prove the universal closure $\forall n \in \omega \exists x(x = \{\ulcorner\varphi\urcorner \mid Tr_d(\ulcorner\varphi\urcorner, n)\})$. This state of affairs, though, is not irreparable. There is a natural candidate to consider for bridging the gap between the individual instances and the universal closure, namely a form of induction. Let φ be a formula of $\mathcal{L}_{\mathrm{ZFC}}$ whose free variables are among $v_1, \ldots, v_m$. Consider the following schema:

$$\forall v_1, \ldots, v_m([\varphi(0) \wedge \forall n \in \omega(\varphi(n) \rightarrow \varphi(n+1))] \rightarrow \forall n \in \omega(\varphi(n))). \quad \text{(A-Ind)}$$

It is easy to see that this induction schema makes it possible to prove also the aforementioned universal closure. That is to say, a new, expanded system of set theory ZFC+A-Ind $\vdash \forall n \in \omega \exists x(x = \{\ulcorner\varphi\urcorner \mid Tr_d(\ulcorner\varphi\urcorner, n)\})$. In this expanded theory it is possible to prove that for every ZFC theorem φ of any degree n it is the case that $Tr_d(\ulcorner\varphi\urcorner, n)$, i.e., that every ZFC theorem is true. Once we have achieved this result, the provability of consistency of ZFC in ZFC+A-Ind easily follows.

The strategy of defining the relations of satisfaction and truth for formulas of restricted complexity can also be pursued further with a view to getting the general relations of satisfaction and truth for $\mathcal{L}_{\mathrm{ZFC}}$. Again, we can turn to the second-order language $\mathcal{L}_{\mathrm{ZFC}_2}$ and our old acquaintance ZFC_2. Without going into any detail, let us just say that once we have the liberty to employ classes, there is no problem to define the relations of satisfaction and truth for the formulas or sentences of the first-order fragment of $\mathcal{L}_{\mathrm{ZFC}_2}$, and prove the consistency of ZFC and all its truths. This can also be done in MK set theory but not in NBG. As we saw in section 4.6, although $\mathcal{L}_{\mathrm{NBG}}$ is rich enough to express the property of truth for $\mathcal{L}_{\mathrm{ZFC}}$, and although NBG passes the test imposed by the convention T, so it can be thought of as

[31] Cf. Kanamori [2006], p. 238.

a truth definition for $\mathcal{L}_{\mathrm{ZFC}}$, NBG cannot prove that the theorems of ZFC are true or that ZFC is consistent. Nevertheless, if a schema of induction along the lines of (A-Ind) is added as a new axiom, the new extended theory obtained in this way becomes strong enough to prove what is required.

This terminates our account of the two broad strategies leading to the definitions of truth for $\mathcal{L}_{\mathrm{ZFC}}$. In principle, it is not necessary to stop here. We could extend our language by adding variables ranging over still higher-order objects such as third-order classes; employing these higher-order objects, we could define truth for $\mathcal{L}_{\mathrm{ZFC}_2}$ as well as prove the (relative) consistency and truth of ZFC_2. We will not, however, pursue this direction. A reason why will emerge in section 4.8.

4.8 Higher-order Objects and Truth

Let us start with a brief summary. We have seen that the language $\mathcal{L}_{\mathrm{ZFC}}$ cannot contain the truth predicate for $\mathcal{L}_{\mathrm{ZFC}}$, and that ZFC cannot involve a theory of truth for $\mathcal{L}_{\mathrm{ZFC}}$. This, of course, can be generalized. No extension of $\mathcal{L}_{\mathrm{ZFC}}$ can contain the general truth predicate, and no consistent extension of ZFC is capable of defining truth for its own language. However, various partial concepts of truth can be defined within ZFC. This fact is important enough in its own right but what is crucial to recognize is that certain partial concepts of truth can be introduced within suitable extensions of ZFC or $\mathcal{L}_{\mathrm{ZFC}}$ in such a way that these partial concepts coincide with the original general concept of truth for $\mathcal{L}_{\mathrm{ZFC}}$. Obtaining the general truth for $\mathcal{L}_{\mathrm{ZFC}}$ requires, as we have shown, an extension of the language by adding variables ranging over classes. Restricted concepts of truth that do not represent properties of truth for the whole language but only for specific fragments—which, nonetheless, have the significant property that they make provable the truth of ZFC—are obtainable without the necessity to enrich the language. It suffices to extend ZFC by additional axioms.

So we have established that it is possible to define truth for sets if we employ classes. To what extent is this outcome favourable? To answer this question will be a chief task of this section.

It has been said that to define truth for the language of sets with the help of classes 'is not a philosophically satisfying resolution, since we encounter the same old difficulties when we attempt to give an explicit definition of truth for the language of class theory' (McGee [1990], p. 76). The definition of truth for $\mathcal{L}_{\mathrm{ZFC}}$ in ZFC_2 is thus no triumphant accomplishment as this effectively just pushes the problem a step away. This broad argument against the meaningfulness of the whole project we have been developing can be further sharpened into challenging the heuristic direction involved. There is no doubt that classes are generally viewed as more obscure entities than sets. If this is so, however, what reason is there to celebrate if we manage

to solve a problem concerning sets at the expense of obtaining virtually the same problem at the level of classes? In what follows, I will attempt to defend the philosophical significance of the attempts to define truth in a manner described in this chapter as well as the acceptability of the talk of classes. The framework in which this defence will be set is the Zermelian conception of set theory depicted in chapter 3. The latter problem will be dealt with first.

The crucial question is: What is the status of classes as second-order objects? In particular, how do classes fit in the conception of the cumulative hierarchy of sets? Where are they localized in the hierarchy, and what role do they play? In a sense, it seems that the universe of sets is self-sustained, and there is no place for classes at all. A remarkable facet of the cumulative hierarchy is that it is essentially a theory of types, although the types are not explicitly presented as types. To perceive the types in the hierarchy, just recall that we start with the empty domain (or, possibly, with a domain containing urelements), and by the repeated application of the power set operation we reach higher and higher layers of sets, each containing all the subsets of the sets of the preceding layer. Thus there are: individuals (the number of individuals may be zero); sets of individuals (or the null set); sets of sets of individuals, etc. The situation is perfectly analogical to the type-theoretic stratification of relations and functions, according to which there are: individuals; relations and functions of individuals; relations and functions of the relations and functions of individuals, etc. However, there is no need to explicitly introduce this kind of types into ZFC since the set-theoretic paradoxes are not derivable from ZFC even without the explicit specification of types, being blocked by a judicious choice of axioms. Besides, when we need to specify the positions sets occupy within the hierarchy, we can use the rank function ρ, definable within ZFC. Thus, as segments correspond to (cumulative) types, ZFC is capable of stratifying the objects it deals with internally, without any additional type-theoretic apparatus.

In section 3.5, we said that (second-order) classes, i.e., objects of type $((\,))$, are to be interpreted as elements of $V_{\alpha+1} = \wp(V_\alpha)$, where V_α is the domain of a model of ZFC_2 (the domain of sets). Yet, if classes are interpreted as elements of $\wp(V_\alpha)$, why should we bother with them at all? What prevents us from taking $V_{\alpha+1}$ as simply another level of sets on top of V_α from the start? A very neat formulation of the criticism of classes as autonomous objects based on this idea can be found in Drake [1974], p. 17. It is worth quoting in full:

> If we consider V to be the universe of all sets, then classes are subcollections of things from V; if we quantify over classes, this implies that we have the collection of all classes to talk about, and the collection of all classes would be exactly the thing we should take as the next level, following all the levels used to make

> up V. In other words, talking about all classes is tantamount to saying that we have not taken all levels, with no end, but we have another one (the level of classes) which we have not used for making sets.

I will attempt to answer these rather disquieting questions by presenting two different ways of thinking about the universe of sets, which are at the same time two different accounts of what it means to say of a set-theoretic sentence that it is true.

Let us assume that it makes sense to view the universe of sets represented by the entire cumulative hierarchy, V, as a unity. Whatever is a set, no matter how large, belongs to it. Of course, as we know, ZFC_2 cannot prove the existence of any set apart from those that belong to the initial segment V_α where α is an inaccessible ordinal. We have seen that if we supply additional axioms, we will be able to prove also the existence of sets belonging to some higher levels of the cumulative hierarchy, and we have sketched arguments that might support the addition of some such axioms. However, it seems clear that there is no way we could actually exhaust the entire universe, i.e., introduce a system of axioms that would get us all the sets there are. The driving force behind adding axioms asserting the existence of larger and larger sets is thus the conviction that, as Maddy [1988], p. 502, puts it, 'the universe of sets is too complex to be exhausted by a handful of operations, in particular by power set and replacement'. It is this conviction of complexity that makes it impossible, by means of a combination of any acceptable set-theoretic operations, to capture the universe of sets as a whole, as a unity. We have found out that V is closed under the operations of power set and replacement, i.e., that any set obtainable using these operations is a member of V. The conviction that the universe is inexhaustibly complex forces us immediately to refuse the idea that the hierarchy might end with the first level indexed by an inaccessible ordinal. That would simply mean that the universe was not that complex after all. So we accept this inaccessible level as an ordinary member of the hierarchy, and we go on. The general principle underlying this line of thought has been called 'reflection'. It maintains that the universe V is inexhaustible and cannot be completely described; therefore, whatever is true of V must be true already of a certain initial segment V_α. The use of the word 'reflection' suggests that the truth in V is always reflected in a particular initial segment of V.[32]

[32] Various forms of the principle of reflection were studied in Lévy [1960*a*] and Lévy [1960*b*]. Perhaps the most common formulation of the reflection principle is the following:

$$\forall\alpha\exists\beta > \alpha\forall x_1, \ldots, x_n \in V_\beta(\varphi(x_1, \ldots, x_n) \leftrightarrow \varphi^{V_\beta}(x_1, \ldots, x_n)). \qquad \text{(P-Ref)}$$

If φ is a formula of $\mathcal{L}_{\mathrm{ZFC}}$ with only the variables $x_1, \ldots, x_n$ free and without abstraction terms, (P-Ref) is provable in ZFC. In fact, it was proved by Lévy [1960*a*] that (P-Ref) is equivalent to the combination of the axiom of infinity and the schema of replacement. Thus ZFC with these two axioms replaced with the schema (P-Ref) is equivalent to standard

Now, what happens if we enrich this picture by classes? As we have said, class variables are supposed to range over all the subcollections of the universe. But this just seems to cause problems. As if the universe of sets were not mysterious enough, now we obtain another level of mystery. To be more precise, the mystery is usually taken as affecting only proper classes, i.e., the classes that can be put into one-to-one correspondence with V. The other, improper classes are made perfectly acceptable since they are identified with sets. Still, what is it that makes proper classes non-sets? As Tait [1998*a*], p. 280, puts it: 'why, when we treat [a proper class] in all other respects as a set, we nevertheless reject it *as* a set'? (Recall that the difficulties surrounding the concept of class cannot be answered by pointing to the set-theoretic paradoxes, i.e., to the fact that letting (proper) classes in among sets leads to an inconsistency. If adoption of a certain kind of objects leads to a contradiction, it only means that such a kind of objects simply cannot exist. The occurrence of a paradox entails merely *that* something cannot exist in a certain way but it does not provide any reason *why*.)

The answers to the questions raised in the course of this section we propose to accept consists in a profound reversal of the overall viewpoint. The key assumption in the whole approach we have just depicted is that the universe of sets, V, can be understood as a unity, as a determinate totality. We have some understanding of the cumulative hierarchy as a whole and of the place of individual sets within it, and on the basis of this understanding we are able to derive particular or general assertions about sets. However, why should we accept that the universe of sets is a well-defined totality, that it is a meaningful object of understanding? The quantifiers of $\mathcal{L}_{\mathrm{ZFC}}$ are supposed to quantify over all sets there are. Yet, that is just a decision we have made; but what does the 'all the sets there are' mean? And how do we enforce this decision? What if it is impossible to distinguish, 'by specified means, the universe from partial universes'?[33]

I will not try to answer these questions. I will rather attempt to suggest an alternative to this top-down approach, based on Zermelian relativism described in section 3.6. The core idea is that the universe of sets should not be understood as one, as a singular totality of all sets whatsoever that is given to us to investigate.[34] To understand what such a rejection involves, let us outline the approach I wish to defend. We speak of different types of objects, of sets, of classes, etc. Introduction of a type into our language opens up a domain of objects of the given type for us. Introduction of type () opens up the domain of sets, of type (()) the domain of classes, etc. ZFC can be understood as a theory aiming at studying the type of sets, ZFC_2 the types of sets and classes. However, to be given a domain of sets is

ZFC. The provability of (P-Ref) in ZFC motivates stronger reflection principles.

[33]Viz. Lévy [1960*b*], p. 1.

[34]What follows is partly inspired by ideas developed in Tait [1998*a*], pp. 279–283.

something different than to be given the entire, inexhaustible universe of sets. Zermelo's driving idea can be construed as follows. Assume that we have been given a domain of sets, and assume that we have accepted ZFC_2 as a theory appropriate for dealing with objects of that domain. How do we then gain some knowledge about this domain itself, and not just objects that are its members? The answer is, as we saw in section 3.6, that we can extend our theory by adding an axiom asserting the existence of an inaccessible ordinal. This step makes it possible to study the initial domain as a set. And this step may be repeated. The crucial thing to realize is that, seen from this point of view, there is no single set-theoretic universe of all sets; there are just different domains. The cumulative hierarchy V as well as the totality of ordinals, Ω, are to be interpreted not as independently determined collections but as relative with respect to different domains. The symbols 'V' and 'Ω' acquire in each domain a different meaning. As Tait puts it, the universe 'is parasitic off domains' (Tait [1998*a*], p. 282).

What consequences follow from this view for truth? In short, it does not make sense any more to think of set-theoretic truth as being 'reflected down' from the universe of sets. Rather, it has to be introduced 'from below', from within a particular system of set theory deemed suitable for dealing with the objects of a given type. As we have seen, truth for a given language cannot be defined within a theory formulated in that language. This discovery forces us to seek other, less direct paths to introduce the property of truth into set theory. We have described some of these approaches. It has turned out that once we introduce the type of classes, the truth for the language of sets becomes definable. In this sense, the theory of classes is a theory of truth. In contrast with the single set-theoretic universe, classes may lose the mystery that might surround them if we recall what we said already in section 3.6: the level of classes can be reinterpreted as another level of sets in an extended domain. This is also the reason why, although there is no technical obstacle to it, it seems considerably less fruitful to carry on and introduce objects of third-order such as third-order classes ((())). If second-order classes belonging to V_α can be reinterpreted as sets in a larger domain $V_{\alpha+1}$, the third-order classes might be reinterpreted as objects of the domain $V_{\alpha+2}$, and so on.

Chapter 5

Carnap: Truth in Syntax

The definition of truth was arrived at, roughly at the same time, by three independent thinkers: Gödel, Tarski and Carnap. Yet, Gödel did not publish his results. Tarski reached the main part of his conception of truth in 1929. He submitted his celebrated monograph on truth (Tarski [1933]) in 1931 but the essay was not published until 1933. For the German edition (Tarski [1935]), Tarski appended a postscript that considerably modified some conclusions of the monograph. Nonetheless, Tarski's work had not been widely known before the German edition of the 'Concept of Truth' and his talks at the congress of Scientific Philosophy in Paris in September 1935, where the ideas he presented aroused a heated controversy.[1]

Carnap's fate was, at least as far as the race for truth is concerned, quite unfortunate. He arrived—with a certain cue from Gödel (discussed in section 5.6) but otherwise quite independently—at a definition of analyticity which is in the essential respects similar to Tarski's definition of (logical) truth. However, Carnap's results were published in 1934, i.e., after Tarski's Polish edition of the monograph on truth. Moreover, Carnap's definition is unnecessarily complex. It is also presented in a rather peculiar manner. What is most significant, though, is the fact that Carnap made the concept of analyticity central to a very specific, in a sense radical philosophical project of logical syntax. To add to the confusion, after meeting with Tarski, Carnap openly "converted" from syntax to semantics. While it is correct to say that Carnap's way of thinking of truth changed considerably under the influence of Tarski, it would be a mistake to read into this change the understanding of the terms 'syntax' and 'semantics' we have now. As we will see in section 5.5, Carnap's conversion consisted in a relatively subtle move.

As a result of the difficulties surrounding Carnap's treatment of analyticity in *The Logical Syntax of Language*, it is not uncommon to come across serious misconceptions proclaimed in connection with Carnap's syntactic pe-

[1] For the depiction of Tarski's participation at the congress, see Feferman and Feferman [2004], 95–98.

riod.[2] The aim of this chapter is to reconstruct what Carnap's definition of analyticity amounts to, and investigate what role it plays within the broader project of logical syntax.[3]

5.1 Analyticity for Language I

In *The Logical Syntax of Language*, Carnap develops in considerable detail two sample languages, called 'Language I' and 'Language II'. The former is to be a 'definite' language which is to realize the 'finitist' or 'constructivist' tendencies, and as such is supposed to appeal to the intuitionist camp within the philosophy of mathematics.[4] When applied to properties, being DEFINITE is to be read as being effectively decidable.[5] This can be extended to functions: a definite function is one that is effectively computable. However, when Carnap says that Language I is definite he does not mean to say that it is a decidable theory in the contemporary sense, i.e., one whose property of being a theorem is effectively decidable. What he means is explained below on p. 116.

Language I is, in effect, an extension of a version of quantifier-free primitive recursive arithmetic (PRA_0). With a certain simplification, the language of Language I, $\mathcal{L}_\mathrm{I}$, may be described as follows. The primitive symbols of $\mathcal{L}_\mathrm{I}$ include an unlimited supply of individual (numerical) variables, the usual connectives of propositional calculus and the symbol for identity '='. In addition, there are these arithmetical primitives:

- the individual constant '0';
- the successor functor '′';
- functors for primitive recursive functions.

If we restrict the vocabulary of $\mathcal{L}_\mathrm{I}$ to the symbols listed above, and if we require that we add a functor for *each* primitive recursive function, we get the standard language of PRA_0. Yet, $\mathcal{L}_\mathrm{I}$ differs from the language of PRA_0 in three respects: firstly, it is not constructed as a language with a fixed vocabulary, so we may add to it other primitive functors or predicates we need besides those mentioned; secondly, Carnap does not make the requirement that $\mathcal{L}_\mathrm{I}$ should include a functor for each primitive recursive function; and thirdly, Carnap also adds bounded quantifiers as primitive symbols. We will not, however, take bounded quantifiers into account since, given that the

[2] For a typical example of the accusation of Carnap of errors he evidently did not commit, see e.g. Etchemendy [1990], pp. 156–157.

[3] Several parts of sections 5.2, 5.4, 5.5 and 5.6 contain ideas already published as Procházka [2006].

[4] Cf. Carnap [1937], §16, p. 46.

[5] Carnap explains 'definite' as follows: a number property is definite if its 'possession or non-possession by any number whatsoever can be determined in a finite number of steps according to a fixed method' (Carnap [1937], §3, p. 11).

variables are numerical, i.e., they range over natural numbers, the bounded quantifiers can be eliminated by means of finite conjunctions and disjunctions. The rules of formation of $\mathcal{L}_{\mathrm{I}}$ are standard, with a notable convention that a formula containing free variables may be asserted with the effect of asserting its universal closure.

Symbols of $\mathcal{L}_{\mathrm{I}}$ are divided into logical and descriptive. The division is given simply by listing the logical symbols. Logical are the standard symbols of propositional logic, '=', the variables, '0' and '′' as well as as well as any defined symbols whose definitions contain only logical symbols. Descriptive are all the predicates and functors (other than '′') which are either primitive or their definitions contain descriptive symbols. An expression, i.e., a complex of symbols, is logical if it does not contain any descriptive symbol.

The deductive system of Language I, $\mathrm{T_I}$, includes the axioms and rules of inference of propositional logic plus the axioms for identity and a rule of variable substitution. On the arithmetical side, there are:

- a recursive definition of the successor function;
- a recursive definition for every primitive recursive function for which there is a functor in $\mathcal{L}_{\mathrm{I}}$;
- the rule of complete (quantifier-free) induction.

The axioms together with the rules of inference of the deductive system are called 'd-rules'. They determine the relation of being DIRECTLY DERIVABLE, i.e., derivable by a single application of a rule of inference. Roughly, a sentence is DERIVABLE if there is a derivation chain, in which all members are directly derivable from the preceding members or from axioms or premises. It is important to note that while the relation of direct derivability in $\mathrm{T_I}$ is effectively decidable (i.e., definite), the relation of derivability simpliciter is not. (In what follows we will use the term 'inference' instead of 'derivability' simpliciter.) This requires a further explanation.

As is well known, $\mathrm{PRA_0}$ is a complete theory, i.e., for every sentence φ of the language of $\mathrm{PRA_0}$, either $\mathrm{PRA_0} \vdash \varphi$ or $\mathrm{PRA_0} \vdash \neg\varphi$. It is also well known that any consistent complete axiomatized theory is decidable, so $\mathrm{PRA_0}$ (which is consistent if PA is) is decidable. Yet the same does not hold of $\mathrm{T_I}$ as $\mathrm{T_I}$ is not complete with respect to the logical part of $\mathcal{L}_{\mathrm{I}}$. The reason for the divergence lies in the fact that according to Carnap, as we have pointed out, we may assert a formula containing free variables as if it were a well-formed sentence, and not merely a schema. This little trick gives $\mathcal{L}_{\mathrm{I}}$ the ability to express unlimited universality, which the language of $\mathrm{PRA_0}$ lacks. At the same time, it makes $\mathrm{T_I}$ an incomplete theory.[6] A consequence of the incompleteness of $\mathrm{T_I}$ is that, given its undecidable inference relation,

[6] In fact, Carnap does not seem to be sure whether the theory $\mathrm{T_I}$ is or is not complete. He merely says that 'the case may arise' that we come across a sentence such that neither it nor its negation is derivable. Cf. Carnap [1937], §14, p. 37.

it cannot be a decidable theory. Language I then should not presumably be called 'definite'. Carnap acknowledges this and explains that the reason why he considers Language I to be definite is that all of its closed sentences, i.e., sentences that contain no free variables, are definite. In other words, $\mathrm{T_I}$ is a complete and decidable theory with respect to a logical sublanguage $\mathcal{L}_\mathrm{I}$ in which formulas with free variables, which make general assertions possible and which are responsible for the incompleteness of $\mathrm{T_I}$, are not regarded to be assertable sentences.

Now if we consider the expressive powers and limits of $\mathcal{L}_\mathrm{I}$, it is clear that the sentence that witnesses the incompleteness of $\mathrm{T_I}$ will involve a logical predicate $\mathfrak{pr}$ of $\mathcal{L}_\mathrm{I}$ such that every individual instance $\mathfrak{pr}(0)$, $\mathfrak{pr}(0')$, $\mathfrak{pr}(0'')$, ... is derivable but the universal conclusion $\mathfrak{pr}(\mathfrak{z})$ is not derivable, neither is its negation. This estimation makes Carnap to introduce another category of rules for $\mathcal{L}_\mathrm{I}$ that would yield the unprovable sentence $\mathfrak{pr}(\mathfrak{z})$. They are the rules of consequence, or 'c-rules'. We will designate such a system of c-rules as '$\mathrm{S_I}$'. Whereas inference always concerns finite sequences of sentences, consequence is a broader, less restricted relation that permits to obtain a conclusion from an infinite class $\mathfrak{K}$ of sentences. There are two rules that make a sentence $\mathfrak{S}$ of $\mathcal{L}_\mathrm{I}$ a DIRECT CONSEQUENCE of a class $\mathfrak{K}$ of sentences. Firstly, $\mathfrak{S}$ is a direct consequence of $\mathfrak{K}$ if $\mathfrak{K}$ is finite and $\mathfrak{S}$ is derivable from the sentences belonging to $\mathfrak{K}$ without the use of the rule of (complete) induction. The purpose of this rule is merely to subsume the relation of inference under the relation of direct consequence; the rule of complete induction is left out to avoid duplication. The second rule is the ω-rule:[7] let $\mathfrak{K}$ be an infinite class containing all the sentences of the form $\mathfrak{pr}(0)$, $\mathfrak{pr}(0')$, $\mathfrak{pr}(0'')$, ... Then $\mathfrak{pr}(\mathfrak{z})$ is a direct consequence:

$$\{\mathfrak{pr}(0), \mathfrak{pr}(0'), \mathfrak{pr}(0''), \ldots\} \models_{\mathrm{direct}} \mathfrak{pr}(\mathfrak{z}) \qquad \text{(R-}\omega\text{)}$$

Let us say that a consequence-series is a finite sequence of classes such that every member of the sequence is in the relation of direct consequence to its predecessor, and the final member is the singleton containing a given sentence $\mathfrak{S}$. $\mathfrak{S}$ is a CONSEQUENCE of $\mathfrak{K}$, $\mathfrak{K} \models \mathfrak{S}$, if and only if there is a consequence-series leading from $\mathfrak{K}$ to $\{\mathfrak{S}\}$. Having the relation of consequence, we can define that a sentence is ANALYTIC if it is a consequence of an empty class of premises, i.e., if it is a consequence by the rules of $\mathrm{S_I}$ alone.

[7] The ω-rule was introduced in 1930 in a lecture given by Hilbert and published as Hilbert [1931]. In fact, Tarski considered this rule already in 1926 but it was Hilbert who triggered a wave of interest in it. Hilbert used the new rule—about which he made the startling claim that it was finitary—to prove Π_1-completeness of PA+ω-rule. Carnap got acquainted with the ω-rule in 1931, and debated it with Gödel who had had an exchange of views on it with Bernays, Hilbert's collaborator. There is some evidence that Carnap considered putting this rule right into the deductive system $\mathrm{T_I}$ of his Language I but in the end decided to stick with the more familiar complete induction. For the details of the whole story, see Buldt [2004].

A sentence is CONTRADICTORY if it has as a consequence every sentence. Sentences that are neither analytic nor contradictory are SYNTHETIC.

The concept of consequence Carnap defines for Language I raises several questions. First, is it extensionally adequate? That is, does it fulfill the role it is supposed to with respect to $\mathcal{L}_{\mathrm{I}}$, namely does it make every logical sentence of $\mathcal{L}_{\mathrm{I}}$ either analytic or contradictory? Carnap shows that it does, i.e., that S_{I} is a complete theory with respect to the logical part of $\mathcal{L}_{\mathrm{I}}$ (viz. theorem 14.3, p. 40). Secondly, can it be generalized to apply to other languages of a more complex logical structure? Tarski [1936*a*], p. 413, charges that it cannot. However, de Rouilhan [2009], pp. 138–140, shows that Carnap's definition of consequence for Language I can be generalized to Language II, whose logical structure is far more complex; so Tarski's charge has been proven wrong. Moreover, de Rouilhan also suggests that Tarski's own definition formulated in the aforementioned paper can be shown to be coextensive, with respect to sentences of $\mathcal{L}_{\mathrm{I}}$, with that of Carnap's.[8] If this is correct, the difference between the account given by Tarski—which paved the way for the standard model-theoretic account of consequence based on the idea of truth preservation—and Carnap's definition is more in the latter's lack of intuitive appeal than in its inadequacy. At the same time, provided that Carnap's relation of consequence for Language I is, indeed, extensionally adequate, it is interesting in its own right since it throws light on the relationship between the classical rule (or axiom) of complete induction and the infinitary ω-rule. Therefore, Coffa's reprimand that the strategy adopted for defining analyticity for Language I is based on 'an incorrect diagnosis' of the problem and, as a result, is the 'least interesting' one,[9] might not be warranted.

Finally, ω-rule is obviously infinitary and, as such, non-effective. How does it come then that Language I is considered definite, and is supposed to appeal to the intuitionists who disapproved of infinitary operations? To understand this, recall that the system of d-rules, which we have called 'T_{I}', is a system of transformation rules *of* Language I. As Language I is an extension of a version of PRA_0, its language $\mathcal{L}_{\mathrm{I}}$ can express all primitive recursive functions, and the system T_{I} can capture (case-by-case prove) all primitive recursive functions.[10] Language I is thus powerful enough to define a system of Gödel numbering. Via some suitable system of arithmetization, it can represent the functions and concepts that are, in Carnap's terminology, definite, i.e., effectively computable and decidable, respectively. Consequently, as the theory T_{I} is a system of definite d-rules, the relations based on these rules are representable in Language I itself (i.e., expressible in $\mathcal{L}_{\mathrm{I}}$ as well as capturable in T_{I}). On the other hand, the theory S_{I} is a system of indefinite

[8] Cf. de Rouilhan [2009], p. 133.

[9] Cf. Coffa [1991], pp. 287–288.

[10] For the definition of the concept of capturing a function or a relation, see footnote 25 on p. 99.

c-rules *for* Language I which are not, in general, representable in Language I. In the light of this distinction, I propose that we should take Language I as a unity consisting of the language $\mathcal{L}_{\mathrm{I}}$ together with the deductive system T_{I}; that is, we should not think of S_{I} as an internal component of Language I but rather as of a separate metatheory formulated in a different language. If we accept this proposal, the definiteness of Language I will no longer be problematic.

What then is the status of the system S_{I} with respect to Language I? To put it briefly, it is a system of rules formulated *for* Language I in a suitable metalanguage. As T_{I} is an incomplete theory, there will be a logical sentence $\mathfrak{S}$ of $\mathcal{L}_{\mathrm{I}}$ such that neither $\mathfrak{S}$ nor $\neg\mathfrak{S}$ is derivable. This does not mean, though, that T_{I} and $\mathfrak{S}$ are unrelated. On the contrary, there will be a particular relation holding between this $\mathfrak{S}$ and T_{I} but it will be one that Language I cannot represent. In order to represent this relation, we need to approach Language I from a different perspective—which, in this particular case, takes on the shape of the theory S_{I}. To conclude, S_{I} is an indefinite theory introduced in the metalanguage with the aim of describing more fully the logical structure of the definite Language I.[11]

5.2 Analyticity for Language II

The indefinite Language II is an expressively and deductively rich simple theory of types which includes Peano arithmetic (PA). What follows is again a simplification which omits some unnecessary features. The TYPES of expressions are defined as follows:

- 0 is the type of numerical expressions;
- if $t_1, \ldots, t_n$ are the types of the n arguments composing the argument expression $\mathfrak{Arg}$, then $(t_1, \ldots, t_n)$ is the type of the predicate $\mathfrak{Pr}$ occurring in a formula $\mathfrak{Pr}(\mathfrak{Arg})$;[12]
- if $t_1, \ldots, t_m$ and t_n are the types of the argument expressions $\mathfrak{Arg}_1$ and $\mathfrak{Arg}_2$, respectively, then $(t_1, \ldots, t_m : t_n)$ is the type of the functor $\mathfrak{Fu}$ occurring in a formula $\mathfrak{Fu}(\mathfrak{Arg}_1) = \mathfrak{Arg}_2$.

Carnap also speaks of a LEVEL of an expression, which is just a natural number: the level of expressions of type 0 is 0. The level of a predicate and a functor is 1 higher than the greatest level of the arguments. For instance, the level of an expression belonging to the type $((0), (0, 0 : 0))$ is 2.

The vocabulary of the language $\mathcal{L}_{\mathrm{II}}$ includes the standard vocabulary of higher-order predicate logic with identity, with unlimited supply of variables

[11] This reading is also supported by what Carnap himself later says in his remarks on the *Logical Syntax* appended to the *Introduction to Semantics*. See Carnap [1942], p. 247.

[12] Carnap uses '$\mathfrak{Pr}$' with the initial capital letter to signify a predicate expression, i.e., a complex possibly composed of several symbols. '$\mathfrak{pr}$' signifies a syntactically-simple predicate. The same distinction applies to '$\mathfrak{Fu}$' and '$\mathfrak{fu}$'.

of each type and unbounded quantifiers. Among the primitive logical symbols of $\mathcal{L}_{\mathrm{II}}$ are also '0' and the successor functor '′'. Additional primitive descriptive constants of suitable types may be introduced as needed. Yet, only the expressions we have just mentioned are considered logical. The formation rules are standard—observing, of course, the type restrictions—but once again with the convention that a formula with free variables can be asserted as an (open) sentence. This decision makes some subsequent definitions of Carnap's rather more complicated; for this reason, we will not follow Carnap and we will accept as sentences only formulas without free variables.

We take the deductive system T_{II} of d-rules to include the standard axioms and rules of inference of (typed) higher-order predicate calculus with identity. Furthermore, T_{II} contains a version of the axiom of choice and the law of extensionality for predicates and functors, according to which coextensive predicates and functors are interchangeable salva veritate. Finally, arithmetical axioms are represented by the recursive definition of the successor function and the axiom of complete induction. Thus the theory T_{II} is, indeed, very strong. Not only that it includes (higher-order) PA but it incorporates a form of (higher-order) set theory (in which sets are represented by predicates).

Still, T_{II} is an incomplete theory with respect to the logical sentences of $\mathcal{L}_{\mathrm{II}}$. In fact, given that it is built on higher-order logic, which is incomplete, this fact is hardly surprising. Nonetheless, this time Carnap is able to present a particular non-demonstrable sentence as a witness of T_{II}'s incompleteness; moreover, this sentence is first-order (first-level, according to Carnap's terminology). As expected, the incompleteness argument is largely based on the technique developed in Gödel [1931].[13] Yet, Carnap introduces one innovation worth mentioning: he proceeds via a so-called 'diagonalization lemma' (also called 'fixed point theorem').[14]

Carnap adopts a system of Gödel numbering, allowing to code symbols and expressions of $\mathcal{L}_{\mathrm{II}}$ as well as sequences of well-formed formulas such as proofs.[15] As the coding system is standard, it is unnecessary to describe it. The derivation of the diagonalization lemma proceeds in two steps.[16] First, we need two definitions. (Note that '$\mathfrak{A}$' designates any expression of $\mathcal{L}_{\mathrm{II}}$.)

[13] Carnap's argument was—due to space restrictions—left out from the original German edition of *Logische Syntax der Sprache* (Carnap [1934*b*]). It was published separately in two articles as Carnap [1934*a*] and Carnap [1935]; it was restored in the English edition (Carnap [1937]).

[14] It was Carnap who first stated it in print but it is not clear whether the lemma may be considered his own invention. It is undoubtedly implicit in Gödel's proof, which, after all, involves its individual instance. Yet, it cannot be ruled out that it was really Carnap who first arrived at the general form of the lemma. In any case, Gödel [1934], p. 63, gives credit to Carnap.

[15] Cf. Carnap [1937], §19, pp. 54–58.

[16] Cf. Carnap [1937], §35, pp. 129–131.

Carnap defines the function $subst(x, s, y)$ in which $x = \ulcorner\mathfrak{A}_1\urcorner$, $y = \ulcorner\mathfrak{A}_2\urcorner$ and $s = \ulcorner\mathfrak{z}\urcorner$. The value of the function *subst* thus is $\ulcorner\mathfrak{A}_1(\mathfrak{z}/\mathfrak{A}_2)\urcorner$, i.e., the Gödel number of an expression resulting from $\mathfrak{A}_1$ by the substitution of $\mathfrak{A}_2$ for the free variable $\mathfrak{z}$. Carnap goes on to define the function $str(n)$ which assigns to a natural number n its Gödel number. Now comes the second step. Having the two functions in hand, Carnap considers a particular value, $subst(x, 3, str(x))$, where x is a numerical variable and 3 is the Gödel number directly assigned to x as the first numerical variable. Let φ be a formula with a single free variable x, and let us see what happens if we substitute $subst(x, 3, str(x))$ for x to obtain $\varphi(subst(x, 3, str(x)))$. Let us designate this formula as ψ. For any φ, the Gödel number $\ulcorner\psi\urcorner$ can be calculated. This number belongs to the range of the numerical variable x, so we may substitute it for x in ψ. Hence we obtain $\varphi(subst(\ulcorner\psi\urcorner, 3, str(\ulcorner\psi\urcorner)))$. If we abbreviate $subst(\ulcorner\psi\urcorner, 3, str(\ulcorner\psi\urcorner))$ as $\ulcorner\mathfrak{S}\urcorner$, we immediately obtain $\varphi(\ulcorner\mathfrak{S}\urcorner)$. However, if we unpack $\mathfrak{S}$, we see that it is equivalent to $\psi(x/\ulcorner\psi\urcorner)$, which is in turn equivalent to $\varphi(subst(\ulcorner\psi\urcorner, 3, str(\ulcorner\psi\urcorner)))$. If we put the equivalences we have established together, we obtain that $\mathfrak{S} \leftrightarrow \varphi(\ulcorner\mathfrak{S}\urcorner)$. But we have accomplished more than this biconditional. We have shown that for an arbitrary formula φ there will be a sentence $\mathfrak{S}$ such that this equivalence holds. Hence we may state the diagonalization lemma as follows:

> For any formula φ of $\mathcal{L}_{\mathrm{II}}$ with one free variable, there is a sentence $\mathfrak{S}$ such that $\mathrm{T_{II}} \vdash \mathfrak{S} \leftrightarrow \varphi(\ulcorner\mathfrak{S}\urcorner)$. (S-Dia)

(The sentence $\mathfrak{S}$ satisfying the lemma is often called the 'fixed point' for φ.) This is a rather powerful result. The incompleteness of $\mathrm{T_{II}}$ follows from it in a single step; moreover, it can be used to derive additional metamathematical theorems.

To see how the incompleteness of $\mathrm{T_{II}}$ is established using (S-Dia), it suffices to take as φ the predicate designating the property of not being demonstrable in $\mathrm{T_{II}}$. This primitive recursive property can be explicitly defined in $\mathrm{T_{II}}$ but, to save us the effort, we will simply assume that we have it available as $\forall r \neg Bew_{\mathrm{T_{II}}}(r, x)$. Now it immediately follows from (S-Dia) that there is a fixed point $\mathfrak{G}$ such that $\mathrm{T_{II}} \vdash \mathfrak{G} \leftrightarrow \forall r \neg Bew_{\mathrm{T_{II}}}(r, \ulcorner\mathfrak{G}\urcorner)$.[17] From this the following general result can be obtained: let $\mathrm{T_{II}}$ be a consistent, (primitive recursively) axiomatized theory that is able to capture primitive recursive functions as functions, and let $\mathfrak{G}$ be a fixed-point for $\forall r \neg Bew_{T_{\mathrm{II}}}(r, x)$. Then $\mathrm{T_{II}} \nvdash \mathfrak{G}$; and if $\mathrm{T_{II}}$ is ω-consistent, $\mathrm{T_{II}} \nvdash \neg\mathfrak{G}$.[18] (Note that there is no complete "core" such as was $\mathrm{PRA_0}$ in Language I, so there is no way that Language II could be considered definite.)

[17] To get an example of such a sentence $\mathfrak{G}$, the easiest way is perhaps to take the Gödel number $g = \ulcorner\forall r \neg Bew_{\mathrm{T_{II}}}(r, subst(x, 3, str(x)))\urcorner$; the desired sample sentence $\mathfrak{G}$ will then be $\forall r \neg Bew_{\mathrm{T_{II}}}(r, subst(g, 3, str(g)))$. We will not present the syntactic argument establishing that both $\mathfrak{G}$ and $\neg\mathfrak{G}$ are unprovable in $\mathrm{T_{II}}$.

[18] See Smith [2007], pp. 175–176, for the proof.

Carnap responds to the incompleteness of the logical part of Language II in the same way as to that of Language I: we need to introduce a system of c-rules, S_{II}, for Language II. However, this time the adopted strategy follows an exactly opposite direction than before. Instead of setting up the c-rules determining the relation of (direct) consequence, and then defining the concept of analyticity, Carnap first defines analyticity and only then, with its help, the relation of consequence. Furthermore, the central part in the system of c-rules is no longer played by the ω-rule but by the notions of valuation and evaluation, which will be described shortly. Why does Carnap dismiss the ω-rule as a basis of the relation of consequence for Language II? There are two reasons. Firstly, the version of the ω-rule Carnap adopts for Language I involves only numerical variables, and it is not applicable to higher-order variables in any straightforward manner,[19] so it is inapplicable in this form to $\mathcal{L}_{II}$. Nevertheless, as we mentioned in section 5.1, it has been suggested by de Rouilhan [2009] that this difficulty is surmountable, and that consequence for Language II can be defined via the ω-rule. Secondly, as we will see below, Carnap's new definition of consequence for Language II entails that the ω-rule in the form given for Language I is logically valid (analytic).

Before outlining Carnap's new path leading to the definition of analyticity, it needs to be pointed out that it is excessively complicated. Together with the auxiliary concepts, it is elaborated in §34*b*–*d* of the *Logical Syntax*, and it spans over 12 pages of rules and definitions.[20] For this reason, what follows is not a full presentation of the route towards analyticity but rather a contracted reconstruction of the basic ideas ideas behind various Carnap's definitions, with considerable modifications with the aim of making the result more compact and more easily accessible.

The first stage of the progress towards analyticity consists in establishing that every sentence of $\mathcal{L}_{II}$ can be converted into a sentence which has a certain basic form, namely it is atomic, it is a quantifier-free sentence formed using the propositional connectives and negation, or it is a quantified sentence in prenex normal form. Carnap calls this transformation process a 'reduction' and the result a 'reduced sentence'. The whole point is simply to reduce the number of different forms sentences can have to a very limited number, for which the subsequent definitions can be provided. Without disclosing any details, we will assume that, as far as non-atomic sentences are concerned, it suffices to consider the forms: $\neg\mathfrak{S}$, $\mathfrak{S}_1 \vee \mathfrak{S}_2$ and $\forall\mathfrak{v}(\varphi(\mathfrak{v}))$, where φ is a quantifier-free matrix.

[19] Viz. the broader discussion of the ω-rule in Carnap [1937], §48, p. 173.

[20] E.g., Tarski [1936*a*], p. 414, when considering Carnap's definition of contradictoriness, complained that 'Carnap's definition of this concept is too complicated and special to be reproduced here without long and troublesome explanations'. Kleene [1939], pp. 83–84, responded by offering his own, much neater and shorter version of the definition of analyticity for Language II, which is presumably equivalent to Carnap's.

Then come the rules of VALUATION. These rules provide for the assignment of values both to free variables and to constants. The ranges of possible values assigned to variables are chosen according to their types:

- the range of values for variables of type 0 is the class of numerals '0', '0′', '0″', ...;
- the range of values for variables of type $(t_1, \ldots, t_n)$ is the power set of the Cartesian product $\mathfrak{B}_1 \times \ldots \times \mathfrak{B}_n$, where $\mathfrak{B}_1, \ldots, \mathfrak{B}_n$ are valuations for $t_1, \ldots, t_n$;
- the range of values for variables of type $(t_1, \ldots, t_m : t_n)$ is the class of all functions from $\mathfrak{B}_1 \times \ldots \times \mathfrak{B}_m$ into $\mathfrak{B}_n$, where $\mathfrak{B}_1, \ldots, \mathfrak{B}_m, \mathfrak{B}_n$ are valuations for $t_1, \ldots, t_m, t_n$.

The two arithmetical constants present among the logical symbols of $\mathcal{L}_{\text{II}}$ are assigned fixed values: the value of '0' is '0' and the value of the functor '′' is the successor function, i.e., a function from a numerical expression $\mathfrak{St}$ to its successor $\mathfrak{St}'$. It is important to note a rather special aspect of Carnap's strategy, namely that the value assigned to an expression $\mathfrak{A}$ is always of the same type as $\mathfrak{A}$ itself.[21] So the value assigned to a numeral $\mathfrak{St}$ is not a natural number but again a numeral; the value of a predicate is not a class of natural numbers but a class of numerals, etc.

Having obtained the ranges of values that are to be assigned to free variables of arbitrary types and to logical constants, it remains to deal with sentences. Yet what values are to be assigned to sentences? Truth values? Propositions? Carnap chooses differently. He picks out two elementary sample sentences, namely the obviously true '$0 = 0$' and the obviously false '$0 \neq 0$'. These "ultimate" sentences—i.e., strings of symbols—are, for Carnap, syntactic replacements of the truth and the falsehood.[22] They represent, so to speak, an irreducible atomic value that is given to us in language and that has to be accepted as evident.

Now come the rules of EVALUATION. Sentences are complex units but some are more complex than other. Carnap chooses to approach first the atomic forms. Note that the rules are so formulated that they apply equally to atomic sentences (containing only constants) and atomic formulas (containing free variables). At this particular moment, we will use 'φ' to stand both for a formula and for a sentence. The rules of evaluation are:

- let $\mathfrak{B}_1$ and $\mathfrak{B}_2$ be the values assigned to $\mathfrak{Arg}$ and $\mathfrak{Pr}$, respectively; then φ of the form $\mathfrak{Pr}(\mathfrak{Arg})$ is assigned the value '$0 = 0$' if $\mathfrak{B}_1 \in \mathfrak{B}_2$; otherwise it is assigned '$0 \neq 0$';
- let $\mathfrak{B}_1$ and $\mathfrak{B}_2$ be the values assigned to $\mathfrak{A}_1$ and $\mathfrak{A}_2$, respectively; then φ of the form $\mathfrak{A}_1 = \mathfrak{A}_2$ is assigned the value '$0 = 0$' if $\mathfrak{B}_1 = \mathfrak{B}_2$; otherwise it is assigned '$0 \neq 0$'.

[21] Cf. Carnap [1937], p. 109.

[22] Cf. Coffa [1991], p. 291, who considers this choice to be another confusing aspect of Carnap's strategy.

With the technique for assigning values to atomic sentences and formulas in hand, we can proceed to more complex sentential forms. The clauses for the basic propositional connectives can be formulated as follows:

- a sentence $\mathfrak{S}$ of the form $\neg\mathfrak{S}_1$ is assigned the value '$0 = 0$' if $\mathfrak{S}_1$ is assigned '$0 \neq 0$'; otherwise it is assigned '$0 \neq 0$';
- a sentence $\mathfrak{S}$ of the form $\mathfrak{S}_1 \vee \mathfrak{S}_2$ is assigned '$0 = 0$' if at least one of the disjuncts is assigned '$0 = 0$'; otherwise it is assigned '$0 \neq 0$'.

The clauses for the remaining connectives can be easily obtained on the basis of those we have just provided. Given the fact that any well-formed sentence or formula of $\mathcal{L}_{\text{II}}$ without quantifiers is finite, the decision what value it is to be assigned can always be reached in a finite number of steps.

It remains to deal with the assignment of values to sentences containing quantifiers. Here 'φ' stands for a quantifier-free matrix:

- a sentence $\mathfrak{S}$ of the form $\forall\mathfrak{v}(\varphi(\mathfrak{v}))$ is assigned the value '$0 = 0$' if for every assignment $\mathfrak{B}$ of a value to the variable $\mathfrak{v}$ φ is assigned '$0 = 0$'; if there is at least one assignment $\mathfrak{B}$ of a value to the variable $\mathfrak{v}$ for which φ is assigned '$0 \neq 0$', $\mathfrak{S}$ is assigned '$0 \neq 0$'.

The clause for the existential quantifier is derived from the clause for the universal quantifier. Taken all together, the clauses listed above enable us to evaluate a sentence of an arbitrary form.

After all these preparatory steps, we are ready to define analyticity. A logical sentence $\mathfrak{S}$ is ANALYTIC if the evaluation leads to the assignment of the value '$0 = 0$' to $\mathfrak{S}$. It is CONTRADICTORY, if the evaluation yields the value '$0 \neq 0$'. A descriptive sentence $\mathfrak{S}_1$ is analytic if the descriptive constants it contains are replaced by universally bound variables of appropriate types, and the resulting logical sentence $\mathfrak{S}_2$ is analytic. It is contradictory if every valuation for descriptive constants occurring in $\mathfrak{S}_1$ leads to the assignment of '$0 \neq 0$' to $\mathfrak{S}_1$. Carnap applies these two properties also to classes of sentences. A sentential class $\mathfrak{K}$ is analytic if all its members are analytic; it is contradictory if at least one of its members is contradictory. (If a sentential class contains descriptive sentences, the assignment of values to the descriptive constants has to be carried out simultaneously throughout the class, i.e., identical constants occurring in different sentences have to be assigned identical values.) A sentence or a sentential class that is neither analytic nor contradictory is SYNTHETIC.

On the basis of these definitions we can introduce the relation of consequence. A sentence $\mathfrak{S}$ is a CONSEQUENCE of the sentential class $\mathfrak{K}$, $\mathfrak{K} \models \mathfrak{S}$, if and only if $\mathfrak{K} \cup \{\neg\mathfrak{S}\}$ is contradictory. Note that this definition of consequence makes the ω-rule superfluous as it is already established that an infinite class $\{\mathfrak{Pr}(0), \mathfrak{Pr}(0'), \mathfrak{Pr}(0''), \ldots\} \models \forall\mathfrak{z}(\mathfrak{Pr}(\mathfrak{z}))$.[23] It follows that the

[23] Cf. Carnap's theorem 34*f*.10, p. 120.

relation of consequence is not compact, i.e., there are sentences that are consequences of an infinite sentential class without being consequences of any of its finite subclasses.[24]

The complex of the rules leading to the definition of analyticity and contradictoriness constitutes a system of c-rules, S_{II}, for $\mathcal{L}_{II}$. Unsurprisingly, the c-rules are indefinite. It is not, in general, effectively decidable whether an arbitrary sentence of $\mathcal{L}_{II}$ is analytic, contradictory or synthetic. Now the system S_{II} is very strong. Carnap is able to show in S_{II} that every axiom of T_{II} is analytic, and that the relation of derivation is an instance of consequence, i.e., that it is analyticity-preserving. Together, this yields that every theorem of T_{II} is analytic, from which it follows that T_{II} is a consistent theory.[25] So Carnap is able to give a relative consistency proof of T_{II}: T_{II} is consistent if S_{II} is. Indeed, Carnap is well aware of the fact that 'the significance of the presented proof of non-contradictoriness must not be over-estimated' as it is carried out in a stronger theory S_{II} which can itself be inconsistent (Carnap [1937], p. 129). Furthermore, one is able to establish in S_{II} that also some sentences that are not demonstrable in T_{II} are analytic, i.e., are consequences of T_{II}. Examples are Gödel's unprovable sentence $\mathfrak{G}$ or the sentence $\forall r \neg Bew_{T_{II}}(r, \ulcorner 0 \neq 0 \urcorner)$, which can be shown to be analytic. (The proofs are essentially the same as those on p. 99 in section 4.5.) In fact, we may generalize: every logical sentence of $\mathcal{L}_{II}$ is either analytic or contradictory.[26] This result can be also expressed by saying that the theory S_{II} is complete with respect to the logical part of $\mathcal{L}_{II}$ (i.e., S_{II} proves every sentence involving only the logical vocabulary of $\mathcal{L}_{II}$ or its negation). This concludes the technical development of the concept of analyticity for Language II.

5.3 Analyticity in General Syntax

How is analyticity introduced in general syntax and what is the role it is intended to play within the broader framework of Carnap's syntactic project? One should not forget that Languages I and II are just examples; it is the discipline of general syntax that is the true goal. Its purpose is to pro-

[24] Cf. Carnap [1937], §34*f*, p. 117.

[25] Cf. Carnap's theorem 34*i*.24, p. 128.

[26] Cf. Carnap's theorem 34*e*.11, p. 116, and its proof. To see that this is really the case, assume that a sentence $\mathfrak{S}$ is logical and synthetic. By means of the rules of reduction and evaluation of sentences containing propositional connectives, it can be transformed into one of these forms: $0 = 0$, $0 \neq 0$ or $Q_1\mathfrak{v}_1 \ldots Q_n\mathfrak{v}_n(\varphi(\mathfrak{v}_1 \ldots, \mathfrak{v}_n))$, where Q is one of the quantifiers and φ is quantifier-free. As the first form is analytic and the second one contradictory, we only need to consider the last form. Now, the rule for the evaluation of sentences containing quantifiers exhausts all the possibilities: for the universal quantifier, either all assignments yield '$0 = 0$', or there is at least one that yields '$0 \neq 0$'. Similarly for the existential quantifier. In either case, the sentence $\mathfrak{S}$ will be assigned a singular value. Hence, despite the assumption, it will not be synthetic.

vide a broad framework for the philosophy of language and mathematics that Carnap pursues. As it is supposed to be a systematic study of *any* formal theory in *any* language, general syntax faces several challenges that particular theories and languages do not face.

Above all, we have seen that in Languages I and II we start with the division of the vocabulary into two distinct categories, i.e., logical and descriptive, and the division is carried out simply by listing the logical terms. Obviously, this approach is no longer feasible if we are supposed to deal with all languages in general. Hence Carnap reverts the procedure. Instead of seeking a relation that would make any sentence belonging to the antecedently delineated logical part of the language determinate, he defines the distinction between the logical and the descriptive on the basis of an antecedently chosen relation of (direct) consequence. This is to say that the relation of direct consequence is taken as the very starting point of the enterprise of general logical syntax. The introduction of the relation requires that there is a system of formation rules determining what series of symbols are well-formed formulas of the given language, $\mathcal{L}_{\mathrm{A}}$. Once this requirement is satisfied, we simply assume that we have got a system of transformation rules by which the relation of direct consequence is determined.[27] That is, no specification is needed. The relation is taken to be quite arbitrary; any relation of direct consequence will do. The transformation rules constitute the familiar c-rules, S_{A}, on the basis of which the relation of consequence simpliciter can be defined: a sentence $\mathfrak{S}$ is a CONSEQUENCE of a sentential class $\mathfrak{K}_1$ if $\mathfrak{S}$ is a member of every class $\mathfrak{K}$ such that (1) $\mathfrak{K}$ is closed under the relation of direct consequence, i.e., it contains all direct consequences of its subclasses; (2) $\mathfrak{K}_1$ is a subset of $\mathfrak{K}$. A sentence is said to be VALID if it is a consequence of an empty class; it is CONTRAVALID if it has every sentence as a consequence. If we restrict ourselves to those among the transformation rules that are concerned exclusively with effectively decidable properties of (classes of) well-formed formulas, we will get a system of the familiar d-rules, T_{A}, determining the relation of direct derivability and inference.

Being valid is not the same as being analytic. The reason is that the transformation rules may also involve other kinds of rules apart from the logico-mathematical ones; e.g., physical rules or various meaning postulates governing the use of individual descriptive concepts. Thus, if we want to preserve the term 'analytic' as meaning 'true in virtue of the rules of logic (and mathematics) alone'—which is, as we have seen, how Carnap defines analyticity for Languages I and II—we need to separate the rules that are logical from those that are not. In general, we need to separate the logical part of $\mathcal{L}_{\mathrm{A}}$ from the extra-logical, descriptive one. How can we achieve this? Carnap's ingenious idea is to look for a syntactic, formally specifiable property that would make it possible to carry out the division without requiring

[27] Cf. Carnap [1937], §46, p. 169.

going beyond mere symbols. The property in question is to be determinate, i.e., to be either valid or contravalid. A slightly modified version of Carnap's definition of the distinction between logical and descriptive expressions can be sketched as follows.[28] Assume, for the sake of simplicity, that a symbol of $\mathcal{L}_{\mathrm{A}}$ is either logical or descriptive irrespective of context. Let $\mathfrak{a}$ be any symbol, and let $\mathfrak{K}$ be the largest class of symbols that fulfills the two following conditions: (1) for every symbol $\mathfrak{a} \in \mathfrak{K}$ there exists a sentence that is composed only of members of $\mathfrak{K}$; (2) every sentence that is composed solely of members of $\mathfrak{K}$ is either valid or contravalid. Now, to guarantee uniqueness, take $\mathfrak{K}_{\mathrm{l}}$ to be the intersection of all the classes satisfying the aforementioned conditions. An expression is LOGICAL if it is composed only of symbols belonging to $\mathfrak{K}_{\mathrm{l}}$. Otherwise it is DESCRIPTIVE. To put it briefly, logical is the vocabulary belonging to the intersection of classes which are such that everything sayable by means of the symbols they contain is determinate.[29] On the other hand, the descriptive vocabulary does not share this property. While some sentences containing descriptive expressions will also be determinate, there will be descriptive sentences that are indeterminate.[30]

With the definition of logicality in hand, we can easily delimit the logical rules as those in which descriptive expressions occur vacuously, i.e., any other descriptive expressions of the same kind may be uniformly substituted for them without ruining the validity of the sentences.[31] A sentence is logically valid or ANALYTIC if its validity follows from the logical transformation rules alone; similarly for logically contravalid sentences, which are said to be CONTRADICTORY. Again, sentences that are neither analytic nor contradictory are SYNTHETIC. Note that valid sentences whose validity follows from other than logical rules are synthetic; similarly for contravalid sentences.

What are the abilities of general syntax to speak about its own linguistic forms? We know that any theory involving a sufficient amount of arithmetic for developing a system of Gödel numbering can arithmetize its own syntax. Let T_{A} be a consistent axiomatizable theory capable of carrying out arithmetization. Then T_{A} can express and prove a great deal about itself,

[28] Carnap [1937], §50, pp. 177–178.

[29] Cf. Creath [1996], p. 258. Carnap's definition is in several respects problematic. Several authors have been able to come up with examples of languages in which expressions that we would want to have in the logical category, such as the existential quantifier, the symbol for identity or numerals, come out as descriptive (e.g., MacLane [1938], p. 174, Creath [1996], pp. 258–260). On the other hand, there are also examples of languages in which some presumably descriptive expressions qualify as logical (Quine [1960], pp. 398–399). For a broader overview of the challenges to Carnap's definition of logicality as well as a proposal of a solution to them see Bonnay [2009].

[30] Despite the fact that Carnap articulates his definitions rather clearly, there persists a misunderstanding. Potter [2000], pp. 268–269, for instance, presents an argument designed to show that Carnap's definition of logicality, implausibly, makes the universal quantifier descriptive. However, Potter fails to recognize that it is not deductive completeness that characterizes the logical rules but a broader property of analyticity or validity.

[31] Carnap [1937], §51, pp. 180–181.

its abilities and limits. Yet it is important to realize that there are several rather different levels on which it can do that. Firstly, (assuming a version of Church's thesis, we may claim that) $\mathrm{T_A}$ can capture (case-by-case prove) by means of formulas of $\mathcal{L}_\mathrm{A}$ all effectively decidable relations and effectively computable functions. To use Carnap's term, $\mathrm{T_A}$ can capture its own definite syntactic properties and relations. Among other things, $\mathrm{T_A}$ is strong enough not merely to describe but actually to capture the entire definite kernel of Language I. However, while the relation of being directly derivable is definite, the relation of being derivable simpliciter is not. $\mathrm{T_A}$ will not be able to prove for any sentence $\mathfrak{S}$ whether the Gödel number $\ulcorner\mathfrak{S}\urcorner$ has the property $\exists r(Bew_{\mathrm{T_A}}(r, x))$. Still, the property of provability or being a theorem is expressible in $\mathcal{L}_\mathrm{A}$, and definable in $\mathrm{T_A}$. So the theory $\mathrm{T_A}$ can capture the definite relations but, moreover, it is also able to define important indefinite properties and relations such as that of theoremhood. This is the second way in which a theory $\mathrm{T_A}$ is able to speak about its own syntax.

Yet what about indefinite syntactic concepts such as analyticity? Can analyticity be defined and employed within the arithmetized syntax representable within $\mathrm{T_A}$? Carnap's answer is negative. After a consideration of Grelling's and Richard's paradoxes and the paradox of the liar, he reaches the conclusion formulated in the theorem 60*c*.1:

> If S is consistent, or, at least, non-contradictory, then '*analytic* (*in S*)' *is indefinable in* S. (Carnap [1937], §60*c*, p. 219; Smeaton's translation.)

This is nothing else than a version of Tarski's theorem of undefinability of truth. The undefinability is not restricted to analyticity; it holds also of other c-concepts such as validity, consequence, content, etc. Closely connected is the theorem stating that the theory $\mathrm{T_A}$, provided it is consistent, cannot prove its own consistency (viz. Carnap's theorem 60*c*.2, a version of Gödel's second incompleteness theorem). So one may conclude that there are syntactic properties and facts that are inevitably beyond the reach of any single consistent recursively axiomatized theory.

Of course, this is nothing surprising, neither for us nor for Carnap. Carnap knew well Gödel's results and it was with a clear understanding of the incompleteness theorems that he set off on his syntactic project. The meticulous separation of inference from consequence, provability from analyticity, and generally d-rules from c-rules was largely driven by the appreciation of the incompleteness phenomena as well as by the vision of developing a philosophical conception that would encompass all of these deficiencies and erect foundations for a deepened philosophical approach to empirical and pure aspects of knowledge.

So we have got the property of analyticity for $\mathcal{L}_\mathrm{A}$, which is just the truth in virtue of the logical rules of $\mathrm{S_A}$, and the property of validity for $\mathcal{L}_\mathrm{A}$, which is the truth in virtue of the transformation rules of $\mathrm{S_A}$ in general.

Before we conclude the whole expository part concerned with the concept of analyticity, let us just add the following. Carnap uses these properties to obtain an explicatum for the pre-theoretical notion of the meaning or the sense of a sentence, which he calls 'content'. The CONTENT of a sentence $\mathfrak{S}$ is the class of non-valid sentences which are consequences of $\mathfrak{S}$. If the given theory involves only logical rules, which is the case of Languages I and II, the content of a sentence is the class of its non-analytic consequences. The concept of content is thus clearly intended to represent the part of the meaning of a sentence that is not determined by the rules constitutive of the given language; in particular, it aims to represent the extra-logical, non-mathematical part of meaning. It immediately follows from this definition that a valid sentence, whose class of non-valid consequences is empty, has no content, while a contravalid sentence, which has as a consequence any sentence whatsoever, has a total content. Obviously, among the sentences that are without content are all the true sentences of logic and mathematics. Note that this is no deep revelation; it simply follows by definition. Recall that it is a characteristic property of logical expressions that the truth value of all sentences composed solely out of them can be determined by employing the transformation rules of S_A alone. Hence the completeness of the logical part of S_A with respect to the logical sentences of $\mathcal{L}_A$ is not something that needs a proof: it is a defining condition imposed on the notion of logicality. To put it slightly differently, to be a logico-mathematical sentence *means* to be a sentence that does not contain any descriptive vocabulary essentially; therefore all logico-mathematical sentences are analytic or contradictory, i.e., either without any content or with total content. Only sentences involving descriptive vocabulary that does not occur vacuously have non-degenerate content, i.e., mean something but not everything.

In conclusion of this section, there is an important point that needs to be emphasized. Gödel [1953/9], p. 339, words an influential objection against Carnap's analytic–synthetic distinction as well as against his division between the logical and the descriptive. As we have seen, Carnap claims that analytic sentences have null content, i.e., they have no non-valid consequences. At the same time, the truth value of synthetic sentences is supposed not to be determined by the rules of logic (and mathematics) alone; it is supposed to reflect extra-linguistic or factual factors. This implies, according to Gödel, that 'the rules of syntax must be demonstrably consistent, since from an inconsistency *every* proposition follows, all factual propositions included' (ibid.). That is, in order to be sure that our distinction captures what it intends to, we are required, says Gödel, to be able to prove the consistency of the given system. However, it is well known that a proof of consistency of any theory meeting some basic requirements can be given only in a stronger theory whose consistency must be assumed without a proof—or, once again, proved in another, yet stronger theory, etc. In general, the requirement of the provability of consistency of a given system

cannot be satisfied. Consequently, it cannot be maintained that analytic sentences are devoid of extra-linguistic consequences.

This argument of Gödel's seriously misrepresents Carnap's position. Let us pass over the question whether it is justifiable to require the provability of consistency rather than just consistency. The substantial assumption behind Gödel's objection is that there exists an absolute, clear-cut dividing line between the factual or extra-linguistic on the one hand and the logical or intra-linguistic on the other. Yet, it is a crucial part of Carnap's strategy that there is no such absolute division. The distinction between the logical (intra-linguistic) and the descriptive (extra-linguistic) is in no way absolute; it does not exist prior to but is established by the definition of logical vocabulary.[32] Understanding what is factual or extra-linguistic presupposes using a language in which we are able to distinguish between the logical and the descriptive vocabulary. That is, there is no realm of empirical facts that are given to us in some direct, non-mediated fashion. The extra-linguistic or factual is simply whatever does not follow from the rules of the linguistic framework itself—and if every sentence of the given language is determined by the rules of the language, then there is simply no factual element whatsoever. Using Carnap's later distinction, we can say that the questions of what is factual or descriptive and what is logical, or what is contentful and what is without content, etc., are all *internal*, and not external questions.[33]

5.4 From Consequence to Inference

We have seen that no consistent recursively axiomatized theory can capture or define all of its own syntactic properties and relations and prove or state certain significant facts about them. How should one react to this discovery? The dominant perspective Carnap assumes to regard the whole issue is that of the vantage point of the syntax-language, i.e., a metatheory formulated in a metalanguage. We will show below that this perspective is not the only one but for the time being we will use it as a framework to depict a concern connected with the treatment of analyticity developed in sections 5.1–5.3. The key idea involved here is that if we cannot exhaust the totality of syntactic features of a given theory from within the theory itself, we have to approach it from a stronger metatheory formulated in a richer metalanguage. Aiming at making our subsequent discussion maximally clear, let us set up the following convention. Let $\mathrm{T_A}$ be an object-theory formulated in an object-language $\mathcal{L}_\mathrm{A}$. Let $\mathrm{T_M}$ be a different theory formulated in another language $\mathcal{L}_\mathrm{M}$. We will say that $\mathrm{T_M}$ is a metatheory for $\mathrm{T_A}$ if it

[32] Cf., in particular, Carnap [1937], §50, p. 177. For a discussion of this important point, see Friedman [1999*b*], pp. 224–225, and Awodey and Carus [2004], pp. 206, 211–213.

[33] This distinction, formulated in Carnap [1950], pp. 206–213, is further discussed in section 5.6.

includes a system of c-rules S_A for T_A. Assume that we are in possession of a metatheory T_M for T_A.

The question we want to ask now is: What is the relationship between consequence for T_A and inference in T_M? Does the consequence relation persist as a unique, fundamentally distinct c-relation that coexists alongside the relation of inference in T_M? To put it differently, does T_M as a system of logic and mathematics involve two irreducibly different modes of connection between sentences, which ultimately give rise to a logic of inference and a different logic of consequence? Or is there a way of bringing these different kinds of connection down to a common ground? In particular, can the relation of consequence for T_A be subsumed under the relation of inference in T_M? What exactly happens with the concepts such as consequence or analyticity, defined for an object theory, on the level of metatheory? This series of questions can be rephrased into a single one: Can the c-concepts for T_A be explicitly defined within the metatheory T_M, i.e., can they be eliminated in T_M? Or are they explicitly undefinable, so that they must be adjoined in the form of a special axiomatic theory, S_A, to the rest of the deductive apparatus of T_M—if T_M is to serve as a metatheory for T_A? Note what is at stake. Basically, if the c-concepts turned out not to be explicitly definable in terms of d-concepts of the metatheory but had to be introduced as undefined primitives governed by additional axioms, we would have to furnish a sufficient justification for a whole new branch of logic.

Carnap's outlook is rather clear. He emphasizes that the relation of inference is more fundamental than that of consequence, and that the method of derivation, or inference, has a clear logical priority. But not only that; consequence should be regarded in a sense as derivative of inference:

> [I]n fact, the method of derivation always remains the fundamental method; every demonstration of the applicability of any term is ultimately based upon a derivation. Even the demonstration of the existence of a consequence-relation—that is to say, the construction of a consequence-series in the object-language—can only be achieved by means of a derivation (a proof) in the syntax-language. (Carnap [1937], §14, p. 39; Smeaton's translation.)

So in order to show in T_M that a sentence of $\mathcal{L}_A$ is a T_A-consequence of a class of sentences, T_M has to be able to *prove* that this is the case. This is to say that within the metatheory the relation of consequence for the object-theory requires an inferential treatment. Naturally, the same holds of other c-concepts. In particular, to show that a sentence of $\mathcal{L}_A$ is analytic, it has to be proved in T_M that it is such. Hence c-concepts for $\mathcal{L}_A$, defined in the metatheory T_M, rather than being treated as concepts sui generis, are to be entangled in the inferential interplay of the metatheory.[34]

[34]This reading of Carnap's intentions is further confirmed by the statement, made eight

The basic outlook is this: There is no special logic of consequence alongside that of inference. Once defined, the relation of consequence, the property of analyticity, etc., become standard pieces of inferential machinery, and the system of c-rules by which they have been introduced gets absorbed in the metatheory, joining the rest of the deductive apparatus.[35] The idea behind the clear-cut separation of c-rules from d-rules is not that these different collections of rules constitute two fundamentally distinct kinds of logical relations but rather that no logico-mathematical theory meeting some basic requirements can be in its entirety exhausted by a single system of transformation rules. To get a full grasp of syntactic properties characteristic of the given theory and to formulate and establish certain facts about it, we need to extend the transformation rules in a particular direction.

However, Carnap does not give us more than a statement of his basic vision; he does not provide the details of how a metatheory could explicitly define c-concepts such as analyticity or consequence for the object-theory. The definitions he offers are informal, formulated in German or English. Yet, if Carnap's vision is correct, they should be formalizable. Moreover, given that the metatheory can arithmetize its own syntax, the c-concepts of the object-theory must render themselves fully to an internal treatment within the metatheory. With the hindsight, it is not hard to see how such a formalization would proceed, especially in the case of Carnap's Language II. On the other hand, a suitable formalization of the system of c-rules for an arbitrary object-theory in general syntax would have to be more or less guessed. For this reason, we will briefly consider merely the question of formalization of the definition of analyticity for Language II, leaving Language I and general syntax aside.

The individual clauses of the informal definition of analyticity for $\mathcal{L}_{\text{II}}$ as they were presented in section 5.2 above can be restated as clauses involving a function f that takes Gödel numbers of sentences of $\mathcal{L}_{\text{II}}$ as arguments and assigns to them either $\ulcorner 0 = 0 \urcorner$ or $\ulcorner 0 \neq 0 \urcorner$ as a value. Roughly, the clauses of section 5.2 become formulas of the following form: $\forall x((x = \ulcorner \ldots \urcorner \wedge f(x) = \ulcorner 0 = 0 \urcorner) \leftrightarrow \ldots)$, in which all the variables—apart from f—possibly

years after the publication of *Logische Syntax* in the appendix to *Introduction to Semantics*, that at that time he would prefer to speak of 'provable' in a related metatheory to speaking of 'analytic' in the object-theory. Viz. Carnap [1942], p. 247.

[35] This was no peculiarity on the part of Carnap. In Procházka [2008], I attempted to show that Tarski's understanding of the "semantic" concept of logical consequence as defined in Tarski [1936*a*] was driven by a similar conception of logic as a discipline occupied with the relation of inference.

Moreover, this interpretation makes it possible to defend Carnap against a serious objection to his conception of analyticity developed in Friedman [1988], pp. 89–94. With some simplification, Friedman claims that the fact that analyticity is not definable by means of the d-concepts of the object-theory but must be supplemented by c-concepts undermines Carnap's position because the methods based on c-concepts go beyond those that are acceptable in logical syntax. However, if we accept that the c-concepts for the object-theory are absorbed in the d-concept of the metatheory, the problem disappears.

occurring in the given clause are bound. Note that the right-hand side of the biconditional will typically contain occurrences of terms designating the values assigned by the assignment function to the Gödel numbers of the components making up the sentence with the number $\ulcorner \ldots \urcorner$. Let '$Conj(f)$' designate the conjunction of the individual clauses for the different forms of sentences, containing just the variable f free. Then we are able to explicitly define analyticity as:

$$Ana(x) \leftrightarrow_{Def.} \exists f(Conj(f) \wedge f(x) = \ulcorner 0 = 0 \urcorner). \qquad \text{(D-Ana)}$$

Alternatively, (D-Ana) can be stated using the universal quantifier and the implication connective. What is crucial to recognize is the fact that, provided that the individual clauses of the conjunction involve the assignment function, the language in which the definition is formulated has to contain a function variable of type $(0 : t)$, where t is the type of the entity assigned to the given Gödel number. The level number of the given variable will always be 1 higher than that of t.

Obviously, if a concept is explicitly definable, it means that it has become eliminable, i.e., it has been established that the concept does not contribute anything new to the deductive power of the theory or the expressive power of the language for which it has been defined. However, is such an explicit definition always possible? Can one approach any object-language from the vantage point of a metatheory capable of such a definition? Does one always have available variables of an appropriate higher type than are the types of the entities over which range the object-language quantifiers? Carnap does not ask this question, so there is no point in turning to him for the answer. However, this issue is investigated in Tarski's monograph on truth (Tarski [1933]). Unfortunately, Tarski's dealing with the matter is in several respects problematic; nevertheless, as the details and intricate implications are not necessary for the line of reasoning we are following, we will merely attempt to get a gist of a possible answer to our question.

Tarski introduces the notion of the order of a language, which is determined by the order of variables it admits. (Tarski's order of a variable is what Carnap would call 'level'.) In the corpus of the monograph, he distinguishes two basic cases. If the orders of variables a language contains have an upper bound, the language is said to be of finite order. If they are unbounded, the language is of infinite order.[36] The main result reached in the corpus of the text is that an explicit definition of truth can be given for any language of finite order but is impossible for any language of infinite order.[37] The reason is clear: if there is no upper bound on the order of variables occurring in the object-language, the object-language already contains variables of all possible orders, and there is no way in which a variable

[36]Cf. Tarski [1933], pp. 219–221.

[37]Viz. Tarski [1933], p. 265, theses A and B.

of a higher order could be introduced into the metalanguage. However, the conclusion arrived at in the postscript, written for the 1935 German edition of the monograph, is strikingly different: truth can be explicitly defined for *any* formalized language without qualification, no matter whether the orders of its variables are bounded or not. How is this possible? What has brought about the change? What has actually changed is the notion of order. Instead of taking the order of a function or relation variable to be a natural number determined simply by the order of the arguments it actually takes, in the postscript it is an ordinal number determined by the highest order of all variables it could *possibly* take. This new notion of order allows us to consider, for instance, function variables taking as their arguments variables of any finite order such that these orders are not bounded by any natural number (finite ordinal). Such a variable will not be of an absolutely infinite order but it will be of quite a particular infinite order, namely ω. Indeed, once the realm of infinity is breached, we can go on to higher and higher orders: $\omega + 1, \ldots, \omega \cdot 2, \ldots, \omega^2, \ldots, \omega^\omega, \ldots$. Now, the sequence of ordinals serving as measures of the orders of languages is uncountable, whereas—at least if we restrict ourselves to countable languages, i.e., languages with countably many symbols—the totality of different variables of the given language will always be only countable. It follows that it is in principle possible to construct, for any given object-language containing variables of arbitrary high orders, a metalanguage with variables of a yet higher order.[38]

That is, a truth definition can be given for any language containing variables of any specific order. Yet, do all variables have a definite order? Tarski's answer is a negative one. There are languages in which variables are taken to range indiscriminately over entities of all orders; such languages are of indefinite order. Tarski cites, as an example of such a language, the language of Zermelo-Fraenkel set theory. Surprising or dubious as this may be, we will not discuss the question whether it is reasonable to take what is usually considered to be a first-order language as a language of indefinite order.[39] The essential point is that, if we allow for languages of indefinite

[38]The phenomenon of the availability of uncountable orders for countable languages is picked out by Gödel in his famous footnote 48a as the "true reason" for the incompleteness of arithmetic:

> [T]he true reason for the incompleteness inherent in all formal systems of mathematics is that the formation of ever higher types can be continued into the transfinite [...], while in any formal system at most denumerably many of them are available. For it can be shown that the undecidable propositions constructed here become decidable whenever appropriate higher types are added (for example, the type ω to the system P). (Gödel [1931], p. 610; van Heijenoort's translation.)

I.e., if we add variables of appropriate higher types, the explicit truth definition for the original language will be available, and the original undecidable proposition will become provable.

[39]We have touched on the issue of types in ZFC already in section 4.8

order, there is no way we could get a metalanguage of a still higher order and define truth for them. Or is there? Astoundingly, Tarski keeps his thesis that truth is definable for any language intact. The only way he is able to achieve this, though, is by introducing one more notion of order of a language. The order of a language containing variables of indefinite order is no longer determined by the order of expressions it contains but by the order of entities whose existence is required for a given system of axioms formulated in this language to be true.[40] For instance, according to this new notion of order, the order of the language of ZFC—with respect to the ZFC axioms—is nothing else than the smallest inaccessible ordinal. As there is no axiomatic system of set theory capable of proving the existence of absolutely all ordinals, there is always a way of obtaining a language whose order is higher than that of any given object-language. Once these two distinct construals of what it means for a language to be of a higher order are joined together, Tarski's aforementioned fundamental result stated in the postscript is immediate: an explicit definition of truth can be constructed for every formalized language.[41]

The new notion of order, Tarski's third in a row, is rather peculiar. First of all, it has no syntactic component. A usual type theory involves a correspondence between the types of expressions of a language on the one hand and the types of entities designated or expressed by these expressions on the other. By contrast, here we deal exclusively with the ontological side of the equation, which is not reflected in the syntax of the language itself but in the axioms of a *theory* formulated in that language. One and the same language can be assigned different orders with respect to different theories. A language of a certain order, say $\mathcal{L}_{\mathrm{T}_1}$, will be capable of formulating just the same sentences as a language of a higher order, $\mathcal{L}_{\mathrm{T}_2}$, which is the language of the theory T_2 in which the property of truth for $\mathcal{L}_{\mathrm{T}_1}$ is to be definable. What changes is the domain of objects over which the variables are taken to range. Obviously, anything a theory such as T_1 proves must be true if the theory itself is to be true but nothing more is required. Therefore, we may take the domain of objects over which the variables of $\mathcal{L}_{\mathrm{T}_1}$ range to consist exclusively of the objects that T_1 proves to exist. The class of all and only such objects will be an entity of a higher order than anything T_1 can prove, hence whatever theory can prove the existence of such a class will be, according to Tarski's third notion of order, formulated in a language of

[40] See Tarski [1983], p. 271.

[41] Viz. Tarski [1983], p. 273, thesis A. Tarski sometimes says that the precondition for the possibility of the definition of truth for an object-language is that the metalanguage must be "essentially richer" than the object-language (e.g., in Tarski [1933], p. 272–273). Two fundamental ways in which a language can be essentially richer are, among others, the two ways in which it can be of a higher order. For an argument that essential richness was originally mainly an informal way of describing the higher-order condition but gradually evolved into a broader concept, see Ray [2005], pp. 434–441.

a higher order than $\mathcal{L}_{\mathrm{T}_1}$.

However, were not the variables $\mathcal{L}_{\mathrm{T}_1}$ supposed to range over entities of all different orders, i.e., over everything there is? How can we simply step in and restrict the range of these variables to the domain of entities that are provable in T_1? Moreover, in order to formulate the restriction, we need a stronger theory such as T_2, which proves the existence of the domain for $\mathcal{L}_{\mathrm{T}_1}$ with respect to T_1. That is, why do we impose such a restriction on the order of $\mathcal{L}_{\mathrm{T}_1}$ if this very act of restricting the range for the variables raises additional ontological requirements? It has been rather convincingly argued by de Rouilhan [1998] that the addition of the postscript to the monograph on truth marks a profound change in Tarski's conception of logic, namely from the universalism—according to which the language of a simple theory of types is to serve as a universal language in which the totality of human knowledge can be formulated—to its brisk abandonment. This change is not explicitly acknowledged but it is hard to see how the third notion of order could be justified without it. On the other hand, the abandonment of universalism squares very well with it: there is no preferred universal standpoint from which one could survey all the entities there are, and there is no universal language of science for whose orders there does not exist any upper bound. What is left are individual languages and theories with their partial assertions about segments of reality. The universal option being excluded, each of these particular languages is eligible for an explicit truth definition.[42]

After this fairly extensive excursus, it is time to return to Carnap. A key idea put forward by Carnap in the *Logical Syntax* is his principle of tolerance, according to which 'everyone is free to choose the rules of his language and thereby his logic in any way he wishes' (Carnap [1963], p. 55). That is, science in general is seen as an open playing field where everyone is permitted to join in with a specific form of logic or a theory formulated in a particular language he or she has constructed. In the light of this philosophical standpoint, Carnap's program is decisively non-universalist. There is no privileged language; no single system may claim a universal dominance. Moreover, not only that none of the competing systems is entitled to universality, but it is an essential requirement that, if a system is to qualify for joining the competition, it must be able to become an object of study of logical syntax, i.e., it must be possible to make its rules explicit, and define its d-concepts and c-concepts. Note that it does not follow that there is a single, universal system of syntax in which one can study all individual pro-

[42] As Feferman [2008], pp. 85–86, emphasizes, though, even after undergoing this decisive shift, Tarski does not assume the position according to which any formalized language is just a calculus, i.e., a language that is not used with meaning but merely studied. So if universalism is characterized by leaning towards the former side of the celebrated distinction: logic as language vs. logic as calculus, introduced in van Heijenoort [1967*b*], there are indications that Tarski remained a logical universalist even after the postscript.

posed scientific systems. This would be indeed a reductio ad absurdum of Carnap's position. What is required is merely that individual competing systems must be subject to syntactic investigations, and any of the competing systems can take up the task in accordance with its own expressive richness and deductive power. Therefore, although it is certainly not the case that analyticity for any language is definable in any metatheory, one seems to be justified to conclude that there must exist a way to define analyticity for any language without exception.

It remains to point out that analyticity does not necessarily have to be defined for a language in its entirety. A partial property of analyticity may be definable within a given language. Take Language II as an example. Recall that it is a simple hierarchy of types. The language $\mathcal{L}_{\mathrm{II}}$ can be stratified into cumulative segments, called by Carnap 'concentric regions' (Carnap [1937], §29, p. 88). Region 1 contains all the symbols of $\mathcal{L}_{\mathrm{II}}$ except for bound occurrences of any variables other than the variables of type 0; but it contains constants and free occurrences of variables of types $(0, \ldots, 0)$ and $(0, \ldots, 0 : 0)$. Region 2 adds bound occurrences of the aforementioned variables of level 1 plus constants and free occurrences of variables of level 2, etc. In general, region n contains constants and free variables up to level n and bound variables up to level $n-1$. To this stratification of the language there correspond theories $\mathrm{T}_{\mathrm{II}_1}$, $\mathrm{T}_{\mathrm{II}_2}$,..., $\mathrm{T}_{\mathrm{II}_n}$. Any property of analyticity restricted to expressions belonging to individual regions is definable within T_{II}, namely in higher regions:

> If we take as our object-language not the whole of Language II but the single concentric regions [...], then for our syntax-language we have no need to go outside the domain of II. It is true that the concept 'analytic in II_n' is not definable for any n in II_n itself as syntax-language, but it is always definable in a more extensive region II_{n+m} (perhaps always in II_{n+1}). Hence every definition of one of the concepts 'analytic in II_n' (for the various n), and also every criterion for 'analytic in II' with respect to a particular sentence of II, is formulable in II as syntax-language. (Carnap [1937], §34*d*, p. 113; Smeaton's translation.)

In fact, as Tarski [1933] shows, Carnap's hunch that analyticity in $\mathcal{L}_{\mathrm{II}_n}$ is definable in $\mathcal{L}_{\mathrm{II}_{n+1}}$ is correct.

In any case, $\mathcal{L}_{\mathrm{II}}$ can be identified with the region $\mathcal{L}_{\mathrm{II}_\omega}$, from which it follows that every single one of the partial concepts of analyticity is definable within the object-theory T_{II}. Moreover, following Tarski's suggestion in the postscript to the monograph on truth, we may expand Carnap's conception of concentric regions by allowing not just natural numbers but a full system of ordinals as their indexes. Then analyticity for $\mathcal{L}_{\mathrm{II}_\omega}$ will be definable in $\mathrm{T}_{\mathrm{II}_{\omega+1}}$ formulated in $\mathcal{L}_{\mathrm{II}_{\omega+1}}$, etc. There is no hint in that direction in the *Logical Syntax* but we may add that if we decided to permit languages that

would be like the language of Language II except for containing also variables of indefinite level, i.e., variables running through all possible levels without a restriction, there would be no obstacle to adjusting the classification of concentric regions so as it reflected the deductive strength of the theories formulated in this language. Analyticity for any individual region $\mathcal{L}_{\mathrm{II}_\alpha}$ would be definable within any theory capable of proving the existence of the class of entities whose existence is provable in $\mathrm{T}_{\mathrm{II}_\alpha}$.

5.5 Analyticity vs. Truth

Let us return to the way analyticity was defined for Language II. For a number of reasons, this definition is highly noteworthy. Above all, it is manifest that once we lift the restriction on the choice of the domain for the valuation of expressions of type 0 to numerals, and once we replace the occurrences of the basic sentences '$0 = 0$' and '$0 \neq 0$' with 'true' and 'false', respectively, we will get nothing else than Tarski's definition of truth for the logical part of $\mathcal{L}_{\mathrm{II}}$. From this point of view, it is clearly visible that a logical sentence is analytic if and only if it is ordinarily true; it is contradictory if it is ordinarily false. At the same time, however, this is exactly the point where the reach of Carnap's definition ends. Recall that, in order for a descriptive sentence to be assigned the value '$0 = 0$', all descriptive constants have to be replaced by (universally bound) variables, and in order for the sentence to be assigned the value '$0 \neq 0$', all valuations of the descriptive constants have to lead to '$0 \neq 0$'. This twin requirement makes, in effect, the descriptive vocabulary in sentences vacuous. Consequently, the resulting concept is that of logical truth and logical falsity for descriptive sentences, and not that of plain truth and plain falsity. Yet, it would be very easy to transform Carnap's definition of analyticity into that of plain truth applicable also to descriptive sentences. Either we could assign permanent values to descriptive constants of $\mathcal{L}_{\mathrm{II}}$, keeping them fixed for all occurrences of the symbols, or we could simply admit all descriptive symbols into the language in which the definition of truth is given. Then we would just need to drop the additional requirements imposed on descriptive sentences, and we would be done. So it is clear that Carnap's definition of analyticity is in a very straightforward manner extensible to a full definition of truth. In this sense, it is essentially the same as Tarski's definition formulated in the monograph on truth. Nevertheless, Carnap does draw a thick division line between analyticity and truth.

This is all the more curious in the light of the fact that, in §60*b* of the *Logical Syntax*,[43] Carnap does, in a sense, outline a theory of plain truth alongside that of analyticity. With a little effort, the following can be extracted from what he says. Assume an object-language $\mathcal{L}_{\mathrm{T}}$ and a

[43] See Carnap [1937], §60*b*, pp. 214–217.

metatheory S formulated in a metalanguage $\mathcal{L}_{\mathrm{S}}$ which contains the syntactic predicates '$\mathfrak{W}$' and '$\mathfrak{F}$'. The rules governing these predicates are to satisfy three basic conditions: first, every sentence of $\mathcal{L}_{\mathrm{T}}$ is either $\mathfrak{W}$ or $\mathfrak{F}$; second, for any sentence $\mathfrak{S}$ of $\mathcal{L}_{\mathrm{T}}$, $\mathfrak{W}(\text{`}\mathfrak{S}\text{'})$ and $\mathfrak{F}(\text{`}\mathfrak{S}\text{'})$ are sentences; and third, the following holds:

$$\mathrm{S} \vdash \mathfrak{W}(\text{`}\mathfrak{S}\text{'}) \leftrightarrow \mathfrak{S}. \qquad \text{(C-T}')$$

Of course, this is nothing else than Tarski's familiar convention T. Carnap then gives a derivation of a contradiction resulting from the assumption that a language can contain its own truth predicate. Thus, measured by Tarski's articulation of the conditions that any truth predicate must satisfy, Carnap's understanding of the concept of plain truth was perfectly adequate. To repeat, the actual definition of analyticity for $\mathcal{L}_{\mathrm{II}}$ can be very easily transformed, as we have seen, into the definition of plain truth; it suffices to lift the additional condition imposed on descriptive sentences, whose only purpose is to make the concept being defined not the plain truth but the truth of descriptive sentences in virtue of logical vocabulary alone.

Still, Carnap obstinately refuses to recognize the link between the property of analyticity and that of truth. In Carnap [1937], §60*b*, p. 216, he unequivocally acknowledges that it *is* possible to develop a theory of truth for an object-language in a metalanguage in a consistent manner. However, he immediately adds:

> A theory of this kind formulated in the manner of a syntax would nevertheless not be a genuine syntax. *For truth and falsehood are not proper syntactical properties*; whether a sentence is true or false cannot generally be seen by its design, that is to say, by the kinds and serial order of its symbols. (Carnap [1937], §60*b*, p. 216; Carnap's emphasis, Smeaton's translation.)[44]

This is the single explicitly stated reason for Carnap's active effort to shun the extension of the concept of analyticity to plain truth. How should we understand it?

[44]This understanding of the concept of truth was characteristic of the whole of Carnap's "syntactic" period. In September 1932 Gödel reported to him in a letter that he was preparing to give a definition of truth in the planned sequel to the incompleteness article (which was never published). Carnap reaction was that Gödel was mixing up two different terms:

> As to terminology: The term 'true' seems to me very unsuitable; in any case, its usage would not be in accord with general linguistic usage. For according to the latter, the sentence 'Vienna has so and so many inhabitants' is of course true, whereas the definition proposed by you surely does not apply to it. Thus one would surely have to say 'logically true' or 'tautological' or 'analytic'. (Gödel [2003], p. 353; Dawson and Goldfarb's translation.)

It was not until 1935 when Carnap met Tarski in Vienna that Carnap grasped the possibility of defining the concept of plain truth. As the story told in Coffa [1991], p. 304, has it, then the scales fell from Carnap's eyes.

An influential answer to this question comes from Coffa, according to whom the chief reason for Carnap's avoidance of ordinary truth was his "verificationist prejudice".[45] The definition of truth is to determine the class of true sentences. It is obvious that no definition of truth along the Tarskian lines can decide straight away which descriptive sentences are true. This can be done at most for analytic sentences. However, there are, in general, two ways of determining classes of objects that share a certain property. One consists in listing the members of such a class, while the other consists in defining a suitable property which selects exactly the objects we would list if we adopted the first method. Indeed, if we were to proceed according to the first method, we would have to decide upon each sentence whether it is true or not before putting it into the class of true sentences. Yet if we followed the latter procedure, we could merely identify a precisely defined property which would—by its structural, or "syntactic" design—guarantee that only true sentences will penetrate into the desired class. In this way, we would be able to define a class of objects but we would still be left entirely in the dark concerning what the elements of this class are.[46] Carnap, thinking of truth only in terms of the former method, which requires the ability to decide which sentences are actually true and which false, failed to recognize the fruitfulness of the latter method. At least, so says Coffa.

Now, in what follows, we will not pursue the question of what was Carnap's true motivation for shutting out the truth definition. We will rather attempt to present a slightly more subtle, perhaps also more charitable picture that might throw some light on the assumptions and consequences of such a decision. The simple fact is, as we will see, that there is no real use for the definition of plain truth within Carnap's broad philosophical project of logical syntax.

The crucial condition for the concept of plain truth to become definable is, as we have explicitly stated, that the metalanguage has to contain either (translations of) the descriptive vocabulary of the object-language or the metatheory must introduce a primitive interpretation function assigning values to the descriptive constants of the object-language. Carnap was well aware of this requirement, and he discussed it in §62 of the *Logical Syntax* (pp. 227–233). To repeat, there is no technical obstacle to empowering the metatheory in the requisite way, and Carnap must have realized this. Let us assume that we have permitted the metalanguage to contain the descriptive vocabulary. What consequences does this step have? The most basic consequence is that the metalanguage can no longer be considered to be syntactic.[47] It has become a fully-fledged language containing synthetic sentences. If $\mathfrak{S}$ is a synthetic sentence of the object-language, the corresponding sen-

[45] Coffa [1991], p. 304.
[46] Cf. Coffa [1977], p. 229.
[47] For a discussion of this point, see also Ricketts [2007], pp. 220–225.

tences $\mathfrak{S}$ and $\mathfrak{W}(\text{`}\mathfrak{S}\text{'})$ of the metalanguage will obviously also be synthetic. If we have not managed to establish the truth of $\mathfrak{S}$ in the object-theory but if we succeed in establishing its truth in the metatheory, it will only be with the help of additional descriptive rules. Nevertheless, the metatheory as a whole will be largely insufficient to determine the truth value of the synthetic sentences of the object-language. In the light of this fact, what do we gain if we permit the descriptive vocabulary into the metalanguage? The synthetic sentences are already present at the level of the object-language, and letting them in the metalanguage does not get us anything that could not have been gained already on the level of the object-language. In this sense, the descriptive vocabulary is just "excess baggage", as Coffa puts it.[48] Yes, it does make the definition of plain truth possible. Nonetheless, although such an achievement is in itself certainly interesting and important, it belongs to a synthetic study of language, and does not seem to contribute with anything fundamentally revealing to the "analytic" project of logical syntax.

Secondly, if the ban on the inclusion of descriptive vocabulary into the metalanguage were lifted, the profound difference between the concept of analyticity and plain truth would vanish. This upshot might perhaps not be undesirable; however, it goes against the very goal of Carnap's philosophical undertaking. In order to see clearly what goes on here, we need to elaborate a little bit upon Carnap's concept of content of a sentence and a particular aspect of the logical–descriptive distinction. It is a well known fact, openly acknowledged by Carnap on several occasions in the *Logical Syntax* as well as elsewhere, that one of the key sources of his inspiration was Wittgenstein's idea, developed in the *Tractatus*, that sentences of logic have no "subject matter", i.e., they do not represent anything.[49] They can be thought of as by-products of the representative capacity of language. The concept of content is developed precisely to capture this fundamental idea. By definition, only synthetic sentences have proper content, while analytic sentences of logic and mathematics are without content. The latter sentences do not say or represent anything; they do not have any meaning at all. They exemplify, in a sense, a degenerate use of language. Their determinate truth value is a clear sign that they do not convey any genuinely new information. If we construct a sufficiently strong metatheory in a sufficiently rich metalanguage containing exclusively logical vocabulary, such a metatheory will be perfectly sufficient to describe the logical syntax of the object-language. That is, it will bring out the logical scaffolding underlying the non-degenerate, contentful use of the object-language. And what is crucial, since such a metatheory will be logical and its sentences either analytic or contradictory, i.e., without content, it will need no additional

[48] Coffa [1991], p. 303.

[49] Cf. Wittgenstein [1922], e.g., 4.0312, 6.124, pp. 68–69, 164–165.

epistemological justification. Its truths will necessarily impose on anyone who understands its rules and its vocabulary.

Thirdly and lastly, Carnap professed a rather special version of logicism. As this issue is rather complex, will just say the following. It is manifest that there is no reduction of mathematic to logic in the *Logical Syntax*, neither there is any attempt to show that mathematics is in some sense inherently logical. The construction of mathematics is thoroughly axiomatic, and it appears to follow the lead of Hilbert rather than of Frege.[50] No reduction is taking place. As Goldfarb puts it, 'Carnap's true original contribution to philosophy of mathematics [is] a version of logicism that does not require the logicist reduction' (Goldfarb [2009], p. 115). It is based on the idea that what suffices is to show that both logic and mathematics can be constructed as analytic. If we construct a sufficiently rich language and build upon it a reasonably strong logical theory, we will automatically obtain a class of logical sentences of this language—which will turn out either analytic or contradictory—and we will find out that the sentences of classical mathematics are among them. Being analytic, mathematics is without content. However, it can be applied contentfully via synthetic (or non-valid) sentences containing descriptive vocabulary. Now, if we decide not to allow descriptive vocabulary into the metalanguage, we will make the metatheory, i.e., the whole logical syntax, a discipline of pure mathematics. So not only that logical syntax is analytic, i.e., it follows from the very ability to apply the rules of the given language; not only that it is without content, i.e., it does not say anything about extra-linguistic facts; it is also purely mathematical, i.e., it can be interpreted as a theory of numbers or sets. The last point gives some plausibility to Carnap's claim that 'pure syntax is [...] nothing more than *combinatorial analysis*, or, in other words, the *geometry* of finite, discrete, serial structures of a particular kind' (Carnap [1937], p. 7; Carnap's emphasis).

In conclusion, it should be clear that the decision not to give a definition of plain truth but to dispose of it as thoroughly as possible in the syntactic investigations is a consequence of Carnap's broader philosophical outlook. Logical syntax is concerned with the meaning of expressions or different modes of connections or relations between them *only insofar as* they are formally representable, and can be treated by purely logical or mathematical means.[51] The decision to make the adjustments necessary for a definition

[50] Cf. Friedman [1988], p. 83.

[51] Cf. especially Carnap [1937], §71, p. 259, where Carnap asks:

> Is it the business of logic to be concerned with the sense of sentences at all [...]? To a certain extent, yes; namely, in so far as the sense and relations of sense permit of being formally represented. Thus, in the syntax, we have represented the formal side of the sense of a sentence by means of the term 'content'; and the formal side of the logical relations between sentences by means of the terms 'consequence', 'compatible', and the like. (Smeaton's

of plain truth does not invalidate or is not incompatible with the project of logical syntax. There is no serious difficulty in developing simultaneously both logical syntax and descriptive syntax as different, supplementary disciplines. However, the concept of plain truth as well as the related concepts such as content, belonging to the latter discipline, do not contribute anything significant to accomplishing Carnap's principal intentions. Perceived from this perspective, Carnap's subsequent shift to espousing the semantic approach towards the philosophy of language and mathematics does not consist in a sudden realization that the plain truth is, after all, definable but rather in a profound change in the broad philosophical outlook that opened up previously unseen ways of exploiting such a definition.

5.6 Syntax and Arbitrary Classes

There still remains a crucial problem that needs to be dealt with. It is connected with Carnap's definition of analyticity for Language II. We have stated that Carnap's approach towards analyticity does not differ from Tarski's conception of (logical) truth in any essential respects. Nevertheless, Tarski's conception is often called 'semantic', and Tarski himself promoted his explication of truth as a 'semantic conception of truth'.[52] How does it come about that Carnap's concept of analyticity and the associated notions such as the relation of consequence, are considered by Carnap to be syntactic, and not semantic? Being syntactic entails, among other things, that in determining the analyticity or contradictoriness of any logical sentence, we do not need to go beyond the syntactic treatment of the given language. However, is this really so? The fact of the matter is that if we recall the details of the definition of analyticity for Language II, there turns out to be a particular point that makes the proclaimed syntactic nature of analyticity rather curious.

It has been well documented[53] that the final shape the treatment of analyticity for $\mathcal{L}_{\text{II}}$ eventually received stems from Carnap's exchange of letters with Gödel that took place in autumn 1932.[54] The crucial shift that Carnap underwent concerned the higher-order quantification. Carnap had originally thought of analyticity substitutionally: a sentence $\forall x(\varphi(x))$ is analytic if the class of the substitution instances $\{\varphi(0), \varphi(0'), \ldots\}$ is analytic. However, the method of substitution leads to problems when higher-order variables are

translation.)

On the other hand, whatever is not formally representable, he goes on, has no place in logical syntax but is a matter of special sciences.

[52] Cf. Tarski [1944], especially pp. 345–346. See also his broad programmatic support for semantics as a reputable scientific discipline in Tarski [1936*b*].

[53] See, e.g., Coffa [1991], pp. 290–291.

[54] See Gödel [2003], especially pp. 346–351. The correspondence between Carnap and Gödel is a valuable source as it contains a lucid expression of Carnap's intentions.

considered. Look at a very simple sentence $\forall F(F(0) \wedge 0 = 0)$. To determine whether this sentence is analytic, we would need to consider all substitution instances of the form $P(0) \wedge 0 = 0$, where P is a predicate constant. Now let us define a predicate constant $P(x)$ as $\forall F(F(x))$ and execute the substitution. It turns out that we can derive the prenex form, only to obtain once again the original sentence $\forall F(F(0) \wedge 0 = 0)$. That is, we have moved in a circle. This problem does not afflict the variables of type 0 because we have clearly determined the values over which they range beforehand. As Gödel suggests, we are free, even at higher levels, to set up a system of rules that would block the circularity or the infinite regress affecting some predicates, and arrive at a well circumscribed list of substitutable predicates. Nevertheless, erection of such a system of additional rules would amount to ramification of the simple theory of types, together with all the difficulties the ramified theory suffers from. For this reason, Gödel argues that one should rather take the quantifier binding a variable of type (t) as ranging over classes of objects of type t, no matter whether they are specifiable by predicates or not.

The rules of valuation detailed on p. 122 evidence that Carnap followed the advice. Recall that we start with a well determined domain of objects of type 0, namely numerals. This domain is by definition countable, and its members are given to us in language as syntactic objects. In the next step, we proceed onto the level 1 of classes of numerals, and we take our quantifiers to range over *all* of them, even over those that are not specifiable in $\mathrm{T_{II}}$. The particular choice of 0-level objects makes the the domain we are investigating syntactic, and the technique of arithmetization of syntax enables the treatment of any syntactic properties or entities whatsoever in terms of numerals—provided only that we have means for representing them via formulas of the given language. Yet we have just said that this condition is not generally satisfied. But then a number of questions immediately forces upon us: How are such non-representable entities given to us? How can we claim that we are doing syntax if we have to go beyond what is representable in our language? In Carnap's words:

> [J]ust as for every language there are numerical properties which are not definable in it [...], so there are also syntactical properties which are not definable in S. [...] Thus the definition [of analyticity] must not be limited to the syntactical properties which are definable in S, but must refer to all syntactical properties whatsoever. But do we not by this means arrive at a Platonic absolutism of ideas, that is, at the conception that the totality of all properties, which is non-denumerable and therefore can never be exhausted by definitions, is something which subsists in itself, independent of all construction and definition? (Carnap [1937], §34*d*, p. 114; Smeaton's translation.)

Thus even if we begin with a well-behaved countable collection of syntactic objects such as numerals, we soon seem to face a tough choice: either to go beyond mere symbols, or to give up the ambition to reach a general criterion of mathematical truth. In any case, a full study of even very simple systems of syntactic objects of a given language requires additional resources that the language in question is incapable of providing.[55]

How does Carnap respond to this problem? The core of his solution to this problem is already contained in his correspondence with Gödel. Carnap asserts:

> The locution " for every valuation ... " that occurs in the definition can still be expressed in a semantics formulated in a definite language, namely by "$(F)(\ldots)$", since a valuation is of course a semantic predicate. This is possible even though in the semantics under consideration not all possible valuations, that is, predicates, can be defined. (Gödel [2003], pp. 354–355; Dawson and Goldfarb's translation.)

To avoid a misunderstanding, note that the term 'semantics' occurring in this quote was a working term for what Carnap later came to call 'syntax'. Thus the core of Carnap's solution may be seen in the claim that the property of being a valuation is (represented by) a *syntactic* predicate, i.e., a defined predicate of the metalanguage. (Recall that valuation is an assignment of values carried out in a metatheory which is definable in the metatheory and expressible by appropriate formulas of the metalanguage in which the definition of analyticity for the object-language is formulated.) Now once we have a general notion of valuation, there is no obstacle to quantifying over all valuations by means of $\forall F(\ldots)$. The intention that we quantify over the totality of arbitrary valuations is assured just by our using the unrestricted universal quantifier. How simple.

But how do we know that we really quantify over all valuations if the metatheory is essentially incapable of capturing or defining them all? How can we make sure that our intention is not left unfulfilled? Here is Carnap's answer (note that 'S' refers to a metalanguage in which the definition of analyticity for an object-language is formulated):

> That this phrase has in the language S the meaning intended is formally established by the fact that the definition of 'analytic in S' is formulated in the wider syntax-language S_2, again in accordance with previous considerations [...], not by substitutions of the $\mathfrak{pr}$ of S, but with the help of valuations. (Carnap [1937], §34*d*, p. 114; Smeaton's translation.)

[55] In his correspondence with Carnap, Gödel said that the definition of analyticity should not be regarded as a clarification of this concept 'since one employs in it the concepts "arbitrary sets", etc., which are just as problematic' (Gödel [2003], pp. 356–657).

So, the questions concerning the range of the quantifiers of the metalanguage get answered in the process of formulation of the definition of analyticity for this metalanguage. This process is, of course, carried out in a higher metatheory, and it is to involve, once again, the totality of valuations for expressions of the metalanguage, articulated by means of $\forall F(\ldots)$ of the metametalanguage. And so on. In general, the quantifiers occurring in any object-languages are to be interpreted objectually as ranging over arbitrary classes; this interpretation is carried out within the definitions of analyticity for the respective object-languages given within the appropriate metatheories.

Carnap's solution can be viewed from at least two distinct angles. From one point of view, Carnap is, rather ingenuously, avoiding a solution rather than providing one. A particular assumption concerning the range of quantifiers of a given language is required to be adopted. How is it adopted? How is it determined that the quantifiers really range over all or arbitrary classes? On each given level n we are said that this is determined on the level $n+1$. So, as Goldfarb [2009], p. 120, puts it, Carnap's position is 'self-supporting at each level'. This does not mean, though, that it is not wanting or vacuous. For, if Carnap so vehemently opposes 'a Platonist absolutism of ideas', and metaphysics in general, what makes his solution preferable or more acceptable? An answer to this objection is to be found in the principle of tolerance: despite the fact that we employ properties in logical syntax that are not specifiable in a particular theory, we are not committed to any metaphysical views concerning entities lying beyond the reach of language. The reason for this is that we are not making any assertion about the ultimate reality. We are simply putting forward a proposal, or setting up a convention, and observing the logical consequences of such an action. The principle of tolerance, by transforming any vigorous assertions into mere proposals, strips any such claims of their force. Perceived from this angle, Carnap does not need to justify his position. By appealing to the principle of tolerance, he has already managed to remove objections to it.[56]

Yet, there is another way of looking at Carnap's solution. At the end of section 5.3, we mentioned the distinction between the internal and external questions introduced in Carnap's article 'Empiricism, Semantics, and Ontology' (Carnap [1950]). External questions, asked prior to setting up a system of rules governing certain concepts or outside such a system, are essentially problematic since they are being raised without a clear idea of how they can be answered. On the other hand, internal questions are asked within an established framework, and the framework determines the way in which they are to be answered. When an external question is raised, there

[56]Cf. Ricketts [2007], p. 219. Goldfarb [2009], p. 120, sees in this the very motivation behind Carnap's principle of tolerance: 'it is the need to have an opening for the view that logical syntax can use notions not specifiable in a particular system without thereby committing itself to Platonism, infinitarism, or the like.'

are, in principle, three different ways of dealing with it. Firstly, it can be reinterpreted as an internal question, i.e., a system of rules can be found or provided in which an answer to this question will get a clear shape. Secondly, it can be reinterpreted as a proposal concerning the construction or adoption of a particular system of rules. Thirdly, it may be rejected as unclear and lacking a sufficient meaning that would permit its being answered in an intelligible manner.

Now, our problem consists in the realization that in order to define analyticity for an object-language, it does not generally suffice to take the stack of predicates that are available in the metalanguage but we need to consider all properties whatsoever, no matter whether specifiable or not. In a sense, we have to go beyond not only the resources available in the given metalanguage but—as this problem persistently reappears at any higher level—beyond all resources available in any language. Once we manage to fix a system of rules, there will always emerge entities lying outside the reach of the rules of such a system. But how are we to we treat such inaccessible entities? How can we survey them or get some definite grasp of them? Note that, in the light of the distinction between the internal and external questions, such questions are external questions par excellence. The basic point of Carnap's solution is then best viewed as an attempt to *internalize* what was originally asked as an external question, i.e., transform it into a more tangible internal question. The definition of analyticity for an object-language requires the ability to take all the entities the object-language can quantify over as a unity, a completed totality, a well-determined class. This is clearly impossible to achieve at the level of the object-language itself; a task like this transcends the powers of such an object-language, and any question regarding these non-specifiable entities is, with respect to the given object-language, clearly external. However, if we introduce a richer metalanguage with a stronger metatheory in which we manage to fulfill the aforementioned requirement, the formerly external questions concerning the object-language will become internal questions of the metalanguage. Within the expanded framework, they will receive a precise meaning. Of course, once these questions are internalized, there will immediately arise new external questions. At no single level we can get a firm grasp of all that there is. On the other hand, all that there is is in principle graspable and can have a precise meaning, regardless of how high we must climb.

Chapter 6

Conclusion

What have we achieved? What conclusions can be drawn from the effort invested into the development of the systems and concepts investigated in chapters 2 to 5? We said in the introduction (chapter 1) that our principal aim would be to investigate the relationship between syntax and semantics. Assuming that we have all the requisite syntactic resources of a given language at our disposal—which, of course, do not allow us to define the general concept of truth for the given language and to develop a theory of meaning for it in general—what exactly is the key step we must make if we are to breach into the meaning-side of that language? What is the crucial element that semantics possesses but syntax lacks? What does the transition from syntax to semantics consist in?

We have investigated in considerable detail (a cumulative version of) Russell's ramified theory of types, Zermelo's second-order set theory and Carnap's logical syntax. We concentrated above all on the way in which truth can be defined in each theory, on the requirements that a truth definition must satisfy and on the consequences that the feasibility of the truth definition has for our understanding of the given system as a whole. In the remaining sections, we will attempt to discern a pattern that characterizes the way truth has been treated in all the systems studied. Then we will return to the question framing our whole journey, namely that of the relationship between syntax and semantics.

6.1 Truth and Partial Truth: a Summary

The fundamental result is, of course, that the property of truth for a given language $\mathcal{L}_{\mathrm{A}}$ is not, on pain of contradiction, definable by a theory formulated in that language; moreover, it is not even expressible by any predicate of $\mathcal{L}_{\mathrm{A}}$. This obstacle is usually surmounted by adopting a different language, $\mathcal{L}_{\mathrm{M}}$, and defining the property of truth for $\mathcal{L}_{\mathrm{A}}$ in this new language $\mathcal{L}_{\mathrm{M}}$. This is easily stated, and relatively easily executed (although the technical details

can be quite complex). However, as we mentioned in the introduction, this comfortable solution of the problem of truth has certain consequences that are not that easy to digest, at least from the philosophical point of view. What have we achieved that might make one look at the things differently?

Let us first consider what an explicit definition of truth amounts to. If $\mathrm{T_M}$ is a theory formulated in the language $\mathcal{L}_\mathrm{M}$ which is able to explicitly define truth for $\mathcal{L}_\mathrm{A}$, then this property of truth can be seen as eliminable. That is, $\mathcal{L}_\mathrm{M}$ does not have to contain any primitive predicate designating this property and $\mathrm{T_M}$ does not have to involve any special axioms governing the use of such a predicate; still, $\mathcal{L}_\mathrm{M}$ will be able to express this property and $\mathrm{T_M}$ will be able to prove a number of facts about it. Now as long as we remain in $\mathcal{L}_\mathrm{M}$ within the bounds of what is provable, i.e., at the level of inference, it can be said that we are engaged merely in syntax (in the wide sense). It follows that, provided that truth for $\mathcal{L}_\mathrm{A}$ is explicitly definable in $\mathrm{T_M}$ without making use of any resources beyond those belonging to the syntax of $\mathrm{T_M}$, the reasoning about the property of truth in $\mathcal{L}_\mathrm{A}$ can be carried out by syntactic means. Thus what initially appears as a semantic enquiry into the meaning-side of $\mathcal{L}_\mathrm{A}$ can be accomplished within the syntax of a metatheory.

We have seen how truth can be explicitly defined in three different theories, and we have found out that such a definition can be reached in several different ways. Firstly, in sections 2.6, 4.6 and 5.4 we saw that in each of the systems it is possible to define truth for the initial language if we gain the ability to quantify over variables of appropriate higher orders. In Russell's theory of types, the constant Val that makes the explicit truth definition (D-$\mathrm{Tr_r}$) possible is always of order 1 higher than is the order of the corresponding propositional function or proposition. This entails that the property Tr_r can hold only of sentences expressing propositions of a restricted order, namely those of order $n-1$ where n is the order of the given relation Val. In $\mathrm{ZFC_2}$ the property of truth for ZFC is definable by means of the relation Sat_c, whose definition (D-$\mathrm{Sat_c}$) requires quantification over proper classes, i.e., objects of a higher order than the objects over which the quantifiers of ZFC range. In Carnap's logical syntax, the general definition of analyticity (D-Ana) involves an assignment function that is of a higher level than is the level of the entity it assigns to a given expression. To conclude, we have been able to establish that, in each of the three theories, truth is explicitly definable using quantification over variables of a higher order. The drawback is, of course, that the property of truth thus defined is only partial: it is not truth for the entire extended language in which the definition is formulated but only for the initial language that does not involve quantification over the added higher-order variables. Total truth for the whole language might be thought of as the limit of the unbounded sequence of the partial concepts of truth.

Secondly, in sections 4.4 and 4.5 we saw that a partial truth definition

for $\mathcal{L}_{\mathrm{ZFC}}$ can be given for any domain that is a set. In this case, the partial property of truth in a set holds of those sentences of $\mathcal{L}_{\mathrm{ZFC}}$ that are true in the given set. Although there is no set large enough to make a definition of total truth for $\mathcal{L}_{\mathrm{ZFC}}$ possible, we can consider very large sets that do not exhaust the universe of sets but may, in some respects or for some purposes, suffice to approximate it. Thus if we assume that there is a segment V_α, where α is a strongly inaccessible ordinal, we obtain all the sets whose existence is provable in ZFC. This will give us the definition of truth which is powerful enough to establish that whatever ZFC can prove is true and that ZFC itself is consistent. The technique of relativization then can make the resulting truth predicate applicable to all sentences of $\mathcal{L}_{\mathrm{ZFC}}$. In section 5.4 we saw that it was this idea that permitted Tarski to assert that truth is definable for any language whatsoever, and that this method is in principle available also in Carnap's syntax.

Thirdly, in section 4.7 we discussed a converse strategy. Rather than adding higher-order variables or axioms asserting the existence of certain very large sets, we can introduce a classification of sentences of the given language based on their complexity, and define partial properties of truth for the restricted classes of sentences. No additional requirements are needed. The total truth is not available without the extension of the language but again it can be thought of as a limit of the unbounded sequence of the partial truth predicates. Naturally, if we extend the language and permit quantification over the variables of a higher-order, the total truth for the initial language will become definable.

Fourthly, in section 5.1 we saw that analyticity for Carnap's Language I is based on the infinitary ω-rule. This approach is quite problematic. Not only that it is not applicable to more complex languages in a straightforward manner but, more importantly, the use of an infinitary rule goes decisively beyond the methods usually permitted in syntax. Nevertheless, aside from these difficulties, it is worth emphasizing that the effect of the ω-rule is in a way comparable to that of the addition of higher-order variables. To see this, take, e.g., PA. The property of truth for $\mathcal{L}_{\mathrm{PA}}$ will be definable in second-order PA_2, the reason being that PA_2 can make use of the quantification over all sets of natural numbers, among which there is also the set consisting exclusively of natural numbers. On the other hand, if we remain within the first-order language $\mathcal{L}_{\mathrm{PA}}$ and instead of enlarging it we add the ω-rule as an additional rule to the deductive apparatus of PA to obtain PA^ω, we will also gain the ability to define truth for $\mathcal{L}_{\mathrm{PA}}$.[1] The way the ω-rule works can be seen in a close analogy with coming into possession of the set of all

[1] As PA^ω is a complete theory with respect to $\mathcal{L}_{\mathrm{PA}}$, truth can be identified with provability. However, PA^ω will no longer be a recursively axiomatizable theory. In particular, valid proofs of PA^ω will not be effectively enumerable. This blocks Gödel's construction of a sentence that is true if and only if it is unprovable. To see how the completeness of PA^ω can be established, see Hazen [1998], p. 514.

natural numbers: to be able to apply the rule, we must be able to attribute a given property to every single natural number. This means, in effect, that we need to complete the list exhausting the natural numbers and not containing anything but natural numbers. This list can be seen in analogy to the set of natural numbers; the rule thus provides, in its own way, the interpretation for the universal quantifier of $\mathcal{L}_{\mathrm{PA}}$. All this is achieved in a theory formulated within one and the same language $\mathcal{L}_{\mathrm{PA}}$. The price paid is that PA^{ω} cannot be viewed as a formal theory, which means that it should not be viewed in any reasonable sense as syntactic. In the light of this, the ω-rule belongs to semantics, and not to syntax.

6.2 Truth, Truths and the Metalanguage

It is time now to draw some consequences. First of all, if we restrict ourselves solely to the resources available in syntax, we have to give up any hope that we can introduce into our language $\mathcal{L}_{\mathrm{M}}$ a total truth predicate, i.e., a truth predicate applicable to any sentence of this language. If this is what has been required from us, we have failed. A totally applicable truth predicate seems to require decisively non-syntactic methods such as application of the ω-rule or some kind of non-mediated, non-linguistic access to the linguistic meaning. So the first conclusion is that, in syntax, we have to abandon the idea of a total property of truth.

All is not lost, though. We can still define partial concepts of truth, suitably restricted so that no contradiction issues. There are two basic options. Firstly, in order to define such a concept of partial truth, it is necessary to impose a clear restriction on the vocabulary or the formation rules of the given language which will determine a particular syntactic region $\mathcal{L}_{\mathrm{A}}$ within the language $\mathcal{L}_{\mathrm{M}}$. The partial truth predicate will be applicable only to the sentences of the given region, not to the remaining sentences of $\mathcal{L}_{\mathrm{M}}$. Of course, as we have seen, not every circumscription leads to a successful definition of partial truth for the given region. It is required either that $\mathcal{L}_{\mathrm{M}}$ be able to employ quantified variables of a higher order than are those of $\mathcal{L}_{\mathrm{A}}$,[2] or that it contain sentences of a higher quantificational complexity. Secondly, we can proceed in terms of theories. Assume that we have a theory T_{M} formulated in $\mathcal{L}_{\mathrm{M}}$ that can prove the existence of the totality w of entities that a theory T_{A} formulated in the same language is capable of proving. Then we can define in T_{M} a partial property of being true in w. Moreover, we can employ the technique of relativization to introduce a partial truth predicate applicable to all sentences of $\mathcal{L}_{\mathrm{M}}$.

[2] We have thus reached, though from quite an opposite direction, a generalization of the assessment formulated and defended in Isaacson [1987], namely that the truths expressible in the language of (first-order) arithmetic that cannot be proved in PA contain "hidden higher-order concepts" (op. cit., pp. 154–155).

In this way, we get a vast number of partial properties of truth, rising upwards in continuous sequences, applicable to specific linguistic regions or with respect to specific sets of entities. The individual concepts of partial truth are comparable with regard to their strength; there will be properties of truth that will be attributable to more sentences than some weaker properties of truth. In any case, the abandonment of the conception of a single total truth for the whole language is replaced by the conception of an unlimited number of partial truths obtained by suitable combination of linguistic restrictions and deductive expansions. This is our second conclusion.

The remaining conclusion is that the significance of the distinction between the object-language and the metalanguage, so vigorously emphasized by both Tarski and Carnap, needs to be reevaluated. Recall that Russell's ramified theory of types was taken to be all formulated in a single language. The individual partial truth predicates were introduced for sentences expressing propositions of particular r-types, according to the r-type of the Val relation. This was all done within one and the same language. Church suggested a straightforward way of transforming the single language of the whole type theory into a hierarchy of different languages that would permit a definition of truth for every member of the hierarchy in its entirety in a language of a higher order. Carnap's imagery of "concentric regions" can be approached in the same way. Either we can view the regions as cumulative levels within one and the same language, namely that of the simple theory of types of Language II, or we can decide to view them as separate languages, forming an unbounded hierarchy. With the language of set theory, it is the other way round. What is standardly considered is the first-order system of ZFC, which is from the very start restricted to the first-order quantification. Second-order ZFC_2 is perceived as a separate theory. There is, nevertheless, no principal objection to considering set theory as formulated within a full-blooded theory of types projected up to arbitrarily high orders. ZFC and ZFC_2 would then be merely representatives of theories formulated within the two lowest fragments of the common language of this all-embracing type theory. Then there would not be the plurality of languages available of which one would be the object-language and another would serve as a metalanguage but only different fragments or regions within a single language, and the individual properties of truth would be applicable solely to the specific fragments for which they were defined. Finally, with regard to the partial truth definitions for the classes of sentences of restricted complexity, it would seem to be an abuse of language to insist on the distinction between the object-language and the metalanguage. We might probably attempt to propose a separation of the formulas of $\mathcal{L}_{\mathrm{ZFC}}$ into different "languages" in such a way that $\mathcal{L}_{\mathrm{ZFC}_0}$ would contain only formulas of complexity 0, $\mathcal{L}_{\mathrm{ZFC}_1}$ formulas of complexity 1, etc. Yet, this would be bordering on the ridiculous—at least if our persistence were driven only by an arduous effort to preserve the object-language-metalanguage distinction.

We do not mean to say that the distinction between the object-language and the metalanguage is fruitless, and that it should be banned once and for all. But what is essential to realize is, first, that this distinction is often quite arbitrary and does not bring in anything essential without which we could not get along. The introduction of this distinction is, in effect, a decision to favour a certain terminology over another. However, secondly, the failure to recognize that the employment or non-employment of this distinction in the development of a theory of truth reflects what essentially is only a terminological decision can lead to an overall misconstrual of what it is that makes the property of truth definable. The picture behind the object-language-metalanguage distinction is roughly that we have a certain language at our disposal, and in order to define truth for this particular language, we need to make a step upwards onto a suitably constructed metalanguage. That is, the picture is typically that of *expansion*. With every metalanguage, we expand the resources we have available, and it is this additional power that makes the definition of truth possible. However, as we have seen, what is as important as a particular form of expansion is the *restriction* of the sentences for which truth is to be defined. The sentences to which the truth predicate is to be applicable need to be restricted to a particular syntactic form. This step is as important as the addition of the quantifiers of a higher-order. Moreover, as we have pointed out, even within first-order languages we can define the properties of partial truth by going downwards, not upwards, i.e., by setting restrictions determining simpler and simpler classes of sentences of the given language.

To repeat, it is granted that the object-language-metalanguage distinction has a perfectly acceptable and meaningful use. In most cases it is completely natural to invoke it when we attempt to define truth. The very fact that we take the object-language as a completed unity entails that we have already restricted the forms of formulas for which the truth is supposed to be defined. Unfortunately, this fact is not usually explicitly stated and recognized as an essential element, and the distinction itself is thought to be revealing some deep fact about the definability of truth.

The reevaluation of the significance of this distinction might have one more interesting effect. One can naturally say: I accept that we are able to define truth for any language provided that we have a suitable metalanguage in hand. However, what if we do not have any such metalanguage available? In particular, what is to serve as a metalanguage for our natural language or for the universal language of science discussed in section 1.4? There clearly cannot exist such a metalanguage for natural language since natural languages do not have sharp boundaries and they incorporate new vocabulary or new rules at will, so it is not acceptable to forcibly close them and forbid them to include another new layer which could then play the role of metalanguage. The universal language of science is also constructed so as it were indefinitely extensible. Seen through the lens of the object-

language-metalanguage distinction, the task of finding a definition of truth for anything as rich as natural language seems thwarted before it can even start. On the other hand, if we dismiss the distinction, we might get a somewhat different and perhaps more optimistic picture.

Let us stick to the universal language of science as it is more tractable. Obviously, we cannot expect to obtain a definition of total truth for the entire language. However, there are two ways in which we can get relatively interesting results. Firstly, as we know, we have a number of different methods that permit us to construct different hierarchies of partial concepts of truth. Admittedly, no single partial property of truth is sufficient in the absolute sense but some of them may be fully sufficient for some specific purposes, for certain particular language regions or fragments. Moreover, when we get to partial truth properties of very high orders or defined with respect to very large sets, we may regard these properties as approximating the total truth. Secondly, we may follow the path suggested by Carnap in his logical syntax. The universal language of science (or, any of the universal languages of science) will have a logico-mathematical part. If this logico-mathematical part is constructed as a simple theory of types similar to Language II, and if it contains enough arithmetic to be able to arithmetize its syntax, the universal language of science will contain as an inherent part a mathematical theory exhibiting its logico-mathematical structure, including the partial c-concepts of analyticity and contradictoriness, logical consequence, content, etc. Now, among other things, this purely mathematical component of the universal language of science will contain a theory of ordinals. As we saw in section 5.4, once we have the ordinals in hand, there is no upper bound on the order of variables we are able to introduce. This means that despite the fact that it does not make sense to assume that the universal language of science can be, as a whole, superseded by a yet more extensive metalanguage, there is no principal obstacle to accepting that, within its logico-mathematical component, there is no bound on the definability of the partial concepts of analyticity. If we translate this idea into the language of the object-language-metalanguage distinction, this means that within the universal language of science, to which no suitable metalanguage exists, there is an unbounded sequence of higher and higher languages that make it possible to define the individual partial properties of analyticity for its logico-mathematical component. As this sequence is unbounded, a metalanguage of this kind is *always* available.

6.3 Semantics and the Absolute

What light does all we have said so far throw on the problem with which we started, namely on the relationship between syntax and semantics? Where does the semantic enterprise start? What is the key move that makes it

possible to introduce semantic notions such as truth or analyticity? We have already identified the two crucial steps on which a truth definition depends: the restriction imposed on the form of sentences to which the truth predicate is to be applicable, and the expansion of the expressive power by making use of additional quantified variables, which typically have to be of a higher-order, or the expansion of the deductive power by adding stronger rules or axioms. The former of these conditions is best viewed as purely syntactic; it merely involves a separation of sentences into classes on the basis of their formal characteristics. Thus only the latter requirement remains as a candidate for the genuinely semantic component of the definition of truth.

This is the ultimate conclusion that this whole thesis tries to make appear plausible. The transition from syntax to semantics consists precisely in gaining the ability to circumvent the totality of objects that an antecedently restricted class of sentences is able to speak of, and to treat this totality as a closed, completed collection. To put it in rather crude terms, in order to gain the ability to define truth and to enter the field of semantics, one must get hold of *everything* the given restricted class of sentences can quantify over, and must be able to treat it as a closed collection. This is not done in any non-mediated, extra-linguistic fashion but by employing suitable quantifiers and variables. However, a further explication is required to avoid a misunderstanding and to point to some consequences of our conclusion.

First of all, a consequence of our investigations is that the notion of "everything" is, to use Russell's term, self-reproductive. In standard languages, the linguistic means which assure that all objects of a given type are taken into account are, indeed, the quantifiers. When we use a language, say $\mathcal{L}_{\mathrm{A}}$, we are confined to the quantified sentences it can formulate, and it is through them that we attempt to grasp the totality of entities we can speak about. Yet, the definition of truth for $\mathcal{L}_{\mathrm{A}}$, say in $\mathcal{L}_{\mathrm{M}}$, requires that $\mathcal{L}_{\mathrm{M}}$ be able to quantify over the totality of entities over which the quantifiers of $\mathcal{L}_{\mathrm{A}}$ run, taken as a completed collection. This does not mean anything else than that the totality of entities that $\mathcal{L}_{\mathrm{M}}$ can quantify over includes entities that were out of reach for $\mathcal{L}_{\mathrm{A}}$, i.e., that transcend its expressive resources. Now we may say that, in general, the totality of entities whose existence can be asserted in language is determined by the linguistic resources available. This assertion surely does not look at all surprising. Still, it is worth emphasizing for the following reason. It contradicts any belief that we can express in language the existence of *absolutely everything*, i.e., that we can grasp the universe in its totality. If truth is defined for a particular language, "everything" is no different, i.e., it is as language-relative as truth is.

With this understanding of the nature of the totality of entities that a language can speak of we may conclude the question of the relationship of syntax and semantics. We have said that manipulation with the quantifiers according to the formation rules of the given language and the rules of inference belongs to syntax in the wider sense. Hence if we possess the

required quantifiers in $\mathcal{L}_{\mathrm{M}}$ that permit us to define truth for $\mathcal{L}_{\mathrm{A}}$, we still remain within the syntactic enterprise. That is, in order to quantify over the completed collection of entities that $\mathcal{L}_{\mathrm{A}}$ can speak of, we do not need to go beyond the syntactic resources of $\mathcal{L}_{\mathrm{M}}$. The whole situation can be expressed in this way: the definition of truth for a restricted language, i.e., the definition of a *semantic* concept, can be given within the *syntax* of an expanded language. Doing semantics essentially involves the recognition that the expressive power of the targeted language or a class of sentences is limited; and this recognition cannot be gained within this very language but can be gained within the syntax of a more powerful language or a more extensive class of sentences. To sum up, once this perspective is adopted, it turns out—to answer our worry from section 1.2—that, to do semantics, there is no need to go outside language to the reality itself. We can carry out meaning analyses and enquire into semantic concepts within language; but, at the same time, this is all we can do.

There is a sense in which the conclusion reached can throw light on the question of what are the entities we investigate in different theories such as natural numbers in arithmetic or sets in set theory. This issue was already discussed in sections 4.8 and 5.6, and the position reached here is in harmony with what we said there. To get a grasp of what the meanings of such theories are, what are the objects they talk about, we need to seize the language of the given theory as a restricted system, and we need to approach it with some additional syntactic resources. This allows us to transform questions that are external with respect to the *restricted system* into questions that are internal with respect to the *expanded system*, which has the resources required for them to be treated. Some systems such as that of set theory are so immensely complex and far-reaching that it is hard to imagine how they could be apprehended in a non-circular fashion in other way than from within suitable extensions of themselves.

To illuminate the point we are trying to make, the following metaphor can perhaps be used. Imagine that we are building a tower which is so vast and so tall that it cannot be surveyed as a whole from any observation point outside the tower. The only way to survey all we have built so far is by adding another floor which will make surveyable all the floors below the top one but not the top floor itself. To survey the whole tower with the top floor included, one must build another floor on top of the former one. And so on, ad infinitum.

This is reminiscent of Kant's solution to the paradoxes (antinomies) of pure reason in transcendental dialectic. The basic point is the idea that spatio-temporal things can be grasped as single objects as well as members of more or less large collections, which can be made larger and larger to approximate the totality of everything that there is. However, the absolute totality of all spatio-temporal things is not within our reach. Importantly, the reason is not that it is so big that it exceeds our cognitive capacities but

rather that the very idea of such a totality stems from a misunderstanding of the way the spatio-temporal things are given to us. That is, the totality in question does not belong to the world of spatio-temporal things but is of a completely dissimilar nature and has a very much different purpose. The idea of the absolute totality is not descriptive; the absolute totality is not something that could be just found or not found to exist. It is a prescriptive or normative unity that has its proper use as a regulative ideal for the cognitive capacity:

> [T]he transcendental ideas are never of constitutive use [...]. [H]owever, they have an excellent and indispensably necessary regulative use, namely that of directing the understanding to a certain goal respecting which the lines of direction of all its rules converge at one point, which, although it is only an idea (*focus imaginarius*)—i.e., a point from which the concepts of the understanding do not really proceed, since it lies entirely outside the bounds of possible experience—nonetheless still serves to obtain for these concepts the greatest unity alongside the greatest extension. (Kant [1998*b*], A797/B825; Guyer and Woods's translation in Kant [1998*a*], p. 591.)

Seen from this viewpoint, it is suggested that the property of total truth for the entire language we use or the absolute totality of everything we can speak of are not concepts or entities on a par with their relative counterparts we can incorporate into the expressive and deductive machinery of language. They are not to be *found*, discovered or shown to exist, they are to be striven for or to be approximated. Hence we are kept in "Wittgenstein's prison" after all.[3] However, the world outside its walls is of a rather different nature than it appears from the inside.

[3]Cf. Awodey and Carus [2007], pp. 33–36.

Bibliography

Aczel, Peter (1988): *Non-Well-Founded Sets*. CSLI Lecture Notes No. 14. CSLI Publications, Stanford. Foreword by John Barwise.

Aspray, William; Kitcher, Philip (eds.) (1988): *History and Philosophy of Modern Mathematics*. Minnesota Studies in the Philosophy of Science, Vol. XI. University of Minnesota Press, Minneapolis.

Awodey, Steve; Carus, A. W. (2004): "How Carnap could have replied to Gödel". In Awodey and Klein [2004], pp. 203–223.

— (2007): "Carnap's dream: Gödel, Wittgenstein, and logical syntax". *Synthese*, vol. 159: pp. 23–45.

Awodey, Steve; Klein, Carsten (eds.) (2004): *Carnap Brought Home. The View from Jena*. Open Court, Chicago.

Balcar, Bohuslav; Štěpánek, Petr (2000): *Teorie množin*. Second, revised edition. Academia, Praha.

Benacerraf, Paul; Putnam, Hilary (eds.) (1983): *Philosophy of Mathematics. Selected Readings*. Second edition. Cambridge University Press, Cambridge.

van Benthem, Johan; Doets, Kees (1983): "Higher-order logic". In Gabbay and Guenthner [1983], pp. 275–329. Reprinted in Gabbay and Guenthner [2001], pp. 189–243. Page references are to the original.

Bernays, Paul (1957): "Betrachtungen zum Paradoxon von Thoralf Skolem". In *Avhandlinger utgitt av Det Norske Videnskaps-Akademie i Oslo*, pp. 3–9.

Bonnay, Denis (2009): "Carnap's criterion of logicality". In Wagner [2009], pp. 147–164.

Boolos, George (1971): "The iterative conception of set". *The Journal of Philosophy*, vol. 68: pp. 215–232. Reprinted in Boolos [1998*a*], pp. 13–29. Page references are to the reprint.

— (1989): "Iteration again". *Philosophical Topics*, vol. 42: pp. 5–21. Reprinted in Boolos [1998*a*], pp. 88–104. Page references are to the reprint.

— (1993): "Whence the contradiction?" *Aristotelian Society Supplementary Volume*, vol. 67: pp. 213–233. Reprinted in Boolos [1998*a*], pp. 220–236.

— (1997): "Constructing Cantorian counterexamples". *The Journal of Philosophical Logic*, vol. 26: pp. 237–239. Reprinted in Boolos [1998*a*], pp. 339–341.

— (1998*a*): *Logic, Logic, and Logic*. Harvard University Press, Cambridge, Massachusetts. Edited by Richard Jeffrey.

— (1998*b*): "Must we believe in set theory?" In Boolos [1998*a*], pp. 120–132. Reprinted in Sher and Tieszen [2000], pp. 257–268, for which it was originally written. Page references are to the first publication.

Buldt, Bernd (2004): "On RC 102-43-14". In Awodey and Klein [2004], pp. 225–246.

Burali-Forti, Cesare (1897): "Una questione sui numeri transfiniti". *Rendiconti del Circolo matematico di Palermo*, vol. 11: pp. 154–164. English translation by Jean van Heijenoort published in van Heijenoort [1967*a*], pp. 105–111.

Butts, R. E.; Hintikka, Jaakko (eds.) (1977): *Logic, Foundations of Mathematics, and Computability Theory*. D. Reidel Publishing, Dordrecht, Netherlands.

Cantor, Georg (1883): *Grundlagen einer allgemeinen Mannigfaltigkeitslehre. Ein mathematisch-philosophischer Versuch in der Lehre des Unendlichen*. Teubner, Leipzig. English translation by William Ewald published in Ewald [1996], pp. 878–920. Page references are to the translation.

— (1891): "Über eine elementare Frage der Mannigfaltigkeitslehre". *Jahresbericht der Deutschen Mathematiker-Vereinigung*, vol. 1: pp. 75–78. English translation by William Ewald published in Ewald [1996], pp. 920–922.

— (1895): "Beiträge zur Begründung der transfiniten Mengenlehre, I." *Mathematische Annalen*, vol. 46: pp. 481–512. English translation by Philip Jourdain published in Cantor [1915], pp. 85–136.

— (1897): "Beiträge zur Begründung der transfiniten Mengenlehre, II." *Mathematische Annalen*, vol. 49: pp. 207–246. English translation by Philip Jourdain published in Cantor [1915], pp. 137–201.

— (1915): *Contributions to the Founding of the Theory of Transfinite Numbers*. Open Court, Chicago. Translated and with an introduction by Philip Jourdain.

Carnap, Rudolf (1934*a*): "Die Antinomien und die Unvollständigkeit der Mathematik". *Monatshefte für Mathematik und Physik*, vol. 41: pp. 263–284.

— (1934*b*): *Logische Syntax der Sprache*. Springer-Verlag, Wien.

— (1935): "Ein Gültigkeitskriterium für die Sätze der klassischen Mathematik". *Monatshefte für Mathematik und Physik*, vol. 42: pp. 163–190.

— (1937): *The Logical Syntax of Language*. Kegan Paul, Trench Trubner and Co., London. English translation by Amethe Smeaton of an expanded version of Carnap [1934*b*].

— (1942): *Introduction to Semantics*. Harvard University Press, Cambridge, Massachusetts.

— (1950): "Empiricism, semantics, and ontology". *Revue Internationale de Philosophie*, vol. 4: pp. 20–40. Reprinted in the supplement to Carnap [1956], pp. 205–221. Page references are to the reprint.

— (1956): *Meaning and Necessity. A Study in Semantics and Modal Logic*. Second edition. University of Chicago Press, Chicago.

— (1963): "Intellectual autobiography". In Schilpp [1963], pp. 1–84.

Chihara, Charles (1973): *Ontology and the Vicious Circle Principle*. Cornell University Press, Ithaca, New York.

Church, Alonzo (1951): "A formulation of the logic of sense and denotation". In Henle *et al.* [1951], pp. 3–24.

— (1976): "Comparison of Russell's resolution of the semantical antinomies with that of Tarski". *The Journal of Symbolic Logic*, vol. 51: pp. 747–760.

Chwistek, Leon (1921): "Antynomie logiki formalnej". *Przeglad Filozoficzny*, vol. 24: pp. 164–171.

Ciesielski, Krzysztof (1997): *Set Theory for the Working Mathematician*. Cambridge University Press, Cambridge.

Cocchiarella, Nino B. (1989): "Russell's theory of logical types and the atomistic hierarchy of sentences". In Savage and Anderson [1989], pp. 41–62.

Coffa, J. Alberto (1977): "Carnap's *Sprachanschauung* circa 1932". In Suppe and Asquith [1977], pp. 205–241.

— (1991): *The Semantic Tradition from Kant to Carnap. To the Vienna Station*. Cambridge University Press, Cambridge.

Cohen, Paul J. (1966): *Set Theory and the Continuum Hypothesis*. W. A. Benjamin, New York. Reprinted by Dover Publications, Mineola, New York, 2008.

Copi, Irving M. (1950): "The inconsistency or redundancy of *Principia Mathematica*". *Philosophy and Phenomenological Research*, vol. 11: pp. 190–199.

— (1958): "The Burali-Forti paradox". *Philosophy of Science*, vol. 25: pp. 281–286.

— (1971): *The Theory of Logical Types*. Routledge and Kegan Paul, London.

Creath, Richard (1996): "Languages without logic". In Giere and Richardson [1996], pp. 251–268.

van Dalen, Dirk; Ebbinghaus, Heinz-Dieter (2000): "Zermelo and the Skolem paradox". *The Bulletin of Symbolic Logic*, vol. 6: pp. 145–161.

Dales, H. G.; Oliveri, G. (eds.) (1998): *Truth in Mathematics*. Oxford University Press, Oxford.

Dauben, Joseph Warren (1979): *Georg Cantor. His Mathematics and Philosophy of the Infinite*. Harvard University Press, Cambridge, Massachusetts.

Davis, Martin (ed.) (1965): *The Undecidable. Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions*. Raven Press, Hewlett, New York.

Devlin, Keith J. (1984): *Constructibility*. Springer-Verlag, Berlin.

Drake, Frank R. (1974): *Set Theory. An Introduction to Large Cardinals*. North-Holland Publishing, Amsterdam.

Dummett, Michael (1963): "The philosophical significance of Gödel's theorem". *Ratio*, vol. 5: pp. 140–155. Reprinted in Dummett [1978], pp. 186–201.

— (1978): *Truth and Other Enigmas*. Harvard University Press, Cambridge, Massachusetts.

— (1981): *Frege. Philosophy of Language*. Second edition. Harvard University Press, Cambridge, Massachusetts.

— (1991): *Frege. Philosophy of Mathematics*. Harvard University Press, Cambridge, Massachusetts.

— (1993*a*): *The Seas of Language*. Clarendon Press, Oxford.

— (1993*b*): "What is mathematics about?" In Dummett [1993*a*], pp. 429–445.

— (1998): "Neo-Fregeans: in bad company?" In Schirn [1998], pp. 369–387.

Ebbinghaus, Heinz-Dieter (2007): *Ernst Zermelo. An Approach to His Life and Work*. Springer-Verlag, Berlin. In cooperation with Volker Peckhaus.

Enderton, Herbert B. (2001): *A Mathematical Introduction to Logic*. Second edition. Academic Press, San Diego, California.

Etchemendy, John (1990): *The Concept of Logical Consequence*. Harvard University Press, Cambridge, Massachusetts.

Ewald, William (ed.) (1996): *From Kant to Hilbert. A Source Book in the Foundations of Mathematics. Volumes 1 & 2*. Oxford University Press, Oxford.

Feferman, Anita Burdman; Feferman, Solomon (2004): *Alfred Tarski. Life and Logic*. Cambridge University Press, Cambridge.

Feferman, Solomon (1999): "Does mathematics need new axioms?" *The American Mathematical Monthly*, vol. 106: pp. 91–111.

— (2008): "Tarski's conceptual analysis of semantical notions". In Patterson [2008], pp. 72–93.

Ferreirós, José (2007): *Labyrinth of Thought. A History of Set Theory and Its Role in Modern Mathematics*. Birkhäuser, Basel. Second, revised edition.

Floyd, Juliet; Kanamori, Akihiro (2006): "How Gödel transformed set theory". *Notices of the American Mathematical Society*, vol. 53: pp. 419–427.

Franzén, Torkel (2005): *Gödel's Theorem. An Incomplete Guide to Its Use and Abuse*. A K Peters, Wellesley, Massachusetts.

Frege, Gottlob (1879): *Begrffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Louis Nebert, Halle. English translation by Stefan Bauer-Mengelberg published in van Heijenoort [1967*a*], pp. 1–82.

— (1884): *Die Grundlagen der Arithmetik. Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Wilhelm Koebner, Breslau.

— (1891): *Funktion und Begriff. Vortrag, gehalten in der Sitzung vom 9. Januar 1891 der Jenaischen Gesellschaft für Medizin und Naturwissenschaft.* Hermann Pohle, Jena. English translation by Peter T. Geach published in Frege [1960], pp. 21–41.

— (1892): "Über Begriff und Gegenstand". *Vierteljahrsschrift für wissenschaftliche Philosophie*, vol. 16: pp. 192–205. English translation by Peter T. Geach published in Frege [1960], pp. 42–55.

— (1893): *Grungesetze der Arithmetik. Begriffsschriftlich abgeleitet. I. Band.* Hermann Pohle, Jena.

— (1903): *Grungesetze der Arithmetik. Begriffsschriftlich abgeleitet. II. Band.* Hermann Pohle, Jena.

— (1960): *Translations from the Philosophical Writings of Gottlob Frege.* Second edition. Basil Blackwell, Oxford. Edited by Peter T. Geach and Max Black.

— (1964): *The Basic Laws of Arithmetic. Exposition of the System.* University of California Press, Berkeley and Los Angeles. Translated and edited by Montgomery Furth.

— (1980): *Philosophical and Mathematical Correspondence.* Blackwell, Oxford. Abridged from the German edition by Brian McGuiness. Translated by Hans Kaal.

Friedman, Michael (1988): "Logical truth and analyticity in Carnap's 'Logical Syntax of Language'". In Aspray and Kitcher [1988], pp. 82–94. Reprinted in Friedman [1999*a*], pp. 165–176.

— (1999*a*): *Reconsidering Logical Positivism.* Cambridge University Press, Cambridge.

— (1999*b*): "Tolarance and analyticity in Carnap's philosophy of mathematics". In Friedman [1999*a*], pp. 198–233.

Friedman, Michael; Creath, Richard (eds.) (2007): *The Cambridge Companion to Carnap.* Cambridge University Press, Cambridge.

Gabbay, Dov M.; Guenthner, Franz (eds.) (1983): *Handbook of Philosophical Logic, Vol. 1. Elemets of Classical Logic.* D. Reidel Publishing, Dordrecht, Netherlands.

— (2001): *Handbook of Philosophical Logic, Second Edition, Vol. 1.* Kluwer Academic Publishers, Dordrecht, Netherlands.

Giere, Ronald N.; Richardson, Alan W. (eds.) (1996): *Origins of Logical Empiricism*. Minnesota Studies in the Philosophy of Science, Vol. XVI. University of Minnesota Press, Minneapolis.

Gödel, Kurt (1931): "Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I". *Monatshefte für Mathematik und Physik*, vol. 38: pp. 173–198. English translation by Jean van Heijenoort published in van Heijenoort [1967*a*], pp. 596–616. Page references are to the translation.

— (1934): "On undecidable propositions of formal mathematical systems". Mimeographed lecture notes, taken by Stephen C. Kleene and J. Barkley Rosser; reprinted with revisions in Davis [1965], pp. 39–74.

— (1944): "Russell's mathematical logic". In Schilpp [1944], pp. 125–153.

— (1947): "What is Cantor's continuum problem?" *The American Mathematical Monthly*, vol. 54: pp. 515–525. A revised and expanded version reprinted in Benacerraf and Putnam [1983], pp. 470–485. Page references are to the reprint.

— (1953/9): "Is mathematics syntax of language?" (Versions III and V). In Gödel [1995], pp. 334–362.

— (1995): *Collected Works, Vol. 3. Unpublished Essays and Lectures*. Oxford University Press, New York and Oxford. Edited by Solomon Feferman.

— (2003): *Collected Works, Vol. 4. Correspondence A–G*. Clarendon Press, Oxford. Edited by Solomon Feferman and John W. Dawson, Jr.

Goldfarb, Warren (1989): "Russell's reasons for ramification". In Savage and Anderson [1989], pp. 24–40.

— (2009): "Carnap's *Syntax* programme and the philosophy of mathematics". In Wagner [2009], pp. 109–120.

Grelling, Kurt; Nelson, Leonard (1907/08): "Bemerkungen zu den Paradoxien von Russell und Burali-Forti". *Abhandlungen der Fries'schen Schule (Neue Serie)*, vol. 2: pp. 300–334.

Griffin, Nicholas (ed.) (2003): *The Cambridge Companion to Bertrand Russell*. Cambridge University Press, Cambridge.

Hallett, Michael (1984): *Cantorian Set Theory and Limitation of Size*. Oxford Logic Guides No. 10. Clarendon Press, Oxford.

— (1996): "Ernst Friedrich Ferdinand Zermelo (1871–1953)". In Ewald [1996], pp. 1208–1218.

Hatcher, William S. (1968): *Foundations of Mathematics*. W. B. Saunders, Philadelphia.

Hausdorff, Felix (1908): "Grundzüge einer Theorie der geordneten Mengen". *Mathematische Annalen*, vol. 65: pp. 435–505.

Hazen, Allen P. (1983): "Predicative logics". In Gabbay and Guenthner [1983], pp. 331–407.

— (1998): "Non-constructive rules of inference". In *Routledge Encyclopedia of Philosophy* (Edward Craig, ed.), pp. 514–517. Routledge, London.

van Heijenoort, Jean (ed.) (1967*a*): *From Frege to Gödel. A Source Book in Mathematical Logic, 1879-1931*. Harvard University Press, Cambridge, Massachusetts.

van Heijenoort, Jean (1967*b*): "Logic as language and logic as calculus". *Synthese*, vol. 17: pp. 324–330.

Henle, Paul; Kallen, Horace M.; Langer, Susanne K. (eds.) (1951): *Structure, Method and Meaning. Essays in Honor of Henry M. Sheffer*. Liberal Arts Press, New York.

Hilbert, David (1900): "Mathematische Probleme. Vortrag, gehalten auf dem internationalen Mathematiker-Kongress zu Paris 1900". In *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen. Mathematisch-Physikalische Klasse*, pp. 253–297. English translation by Mary Winston Newson published in *Bulletin of the American Mathematical Society*, vol. 8 (July 1902): pp. 437–479.

— (1926): "Über das Unendliche". *Mathematische Annalen*, vol. 95: pp. 161–190. English translation by Stefan Bauer-Mengelberg published in van Heijenoort [1967*a*], pp. 367–392. Page references are to the translation.

— (1931): "Die Grundlegung der elementaren Zahlenlehre". *Mathematische Annalen*, vol. 104: pp. 485–494.

Hintikka, Jaakko (ed.) (1969): *The Philosophy of Mathematics*. Oxford University Press, London.

Hodges, Wilfrid (1997): *A Shorter Model Theory*. Cambridge University Press, Cambridge.

Hylton, Peter (1990): *Russell, Idealism, and the Emergence of Analytic Philosophy*. Oxford University Press, Oxford.

Isaacson, Daniel (1987): "Arithmetical truth and hidden higher-order concepts". In Paris Logic Group [1987], pp. 147–169.

Jané, Ignacio (2001): "Reflections on Skolem's relativity of set-theoretical concepts". *Philosophia Mathematica*, vol. 9: pp. 129–153.

Jech, Thomas (ed.) (1974): *Axiomatic Set Theory II. Proceedings of Symposia in Pure Mathematics Vol. 13*. American Mathematical Society, Providence, Rhode Island.

Jech, Thomas (1991): "The infinite". In *Jahrbuch der Kurt-Gödel-Gesellschaft*, pp. 36–44. Wien.

— (2003): *Set Theory*. The Third Millenium edition, revised and expanded. Springer-Verlag, Berlin.

Jung, Darryl (1999): "Russell, presupposition, and the vicious-circle principle". *Notre Dame Journal of Formal Logic*, vol. 40: pp. 55–80.

Kamareddine, Fairouz; Laan, Twan; Nederpelt, Rob (2002): "Types in logic and mathematics before 1940". *The Bulletin of Symbolic Logic*, vol. 8: pp. 185–245.

Kanamori, Akihiro (1996): "The mathematical development of set theory from Cantor to Cohen". *The Bulletin of Symbolic Logic*, vol. 2: pp. 1–71.

— (1997): "The mathematical import of Zermelo's well-ordering theorem". *The Bulletin of Symbolic Logic*, vol. 3: pp. 281–311.

— (2003): *The Higher Infinite. Large Cardinals in Set Theory from Their Beginnings*. Second edition. Springer-Verlag, Berlin.

— (2004): "Zermelo and set theory". *The Bulletin of Symbolic Logic*, vol. 10: pp. 487–553.

— (2006): "Levy and set theory". *Annals of Pure and Applied Logic*, vol. 140: pp. 233–252.

Kant, Immanuel (1998*a*): *Critique of Pure Reason*. Cambridge University Press. Edited and translated by Paul Guyer and Allen W. Wood.

— (1998*b*): *Kritik der reinen Vernunft*. Felix Meiner Verlag, Hamburg. Edited by Jens Timmermann.

Ketland, Jeffrey (1999): "Deflationism and Tarski's paradise". *Mind, New Series*, vol. 108: pp. 69–94.

Kleene, Stephen Cole (1939): "Review of Rudolf Carnap, *The Logical Syntax of Language*". *The Journal of Symbolic Logic*, vol. 4: pp. 82–87.

— (1943): "Recursive predicates and quantifiers". *Transactions of the American Mathematical Society*, vol. 53: pp. 41–73.

Klement, Kevin C. (2001): "Russell's paradox in Appendix B of the *Principles of Mathematics*: Was Frege's response adequate?" *History and Philosophy of Logic*, vol. 22: pp. 13–28.

Kolman, Vojtěch (2002): *Logika Gottloba Frega.* Filosofia, Praha.

Kolman, Vojtěch (ed.) (2006): *From Truth to Proof.* Miscellanea Logica, Vol. VI. Univerzita Karlova v Praze, Praha.

Kolman, Vojtěch (2008): *Filosofie čísla. Základy logiky a aritmetiky v zrcadle analytické filosofie.* Filosofia, Praha.

Kreisel, Georg (1967): "Informal rigour and completeness proofs". In Lakatos [1967], pp. 138–157. Reprinted, with a postscript but without Section 3, in Hintikka [1969], pp. 78–94. Page references are to the original.

Kreisel, Georg; Krivine, Jean-Louis (1966): *Éléments de logique mathématique. Théorie des modèles.* Dunod, Paris.

Kuratowski, Kazimierz (1925): "Sur l'état actuel de l'axiomatique de la théorie des ensembles". *Annales de la Société Polonaise de Mathématique*, vol. 3: pp. 146–147.

Lakatos, Imre (ed.) (1967): *Problems in the Philosophy of Mathematics. Proceedings of the International Colloquium in the Philosophy of Science, London, 1965, Vol. 1.* North-Holland Publishing, Amsterdam.

Landini, Gregory (1992): "Russell to Frege, 24 May 1903: 'I believe I have discovered that classes are entirely superfluous'". *Russell: the Journal of Bertrand Russell Studies*, vol. 12: pp. 160–185.

Lavine, Shaughan (1994): *Understanding the Infinite.* Harvard University Press, Cambridge, Massachusetts.

Lévy, Azriel (1959): "A hierarchy of formulae of set theory (abstract)". *Notices of the American Mathematical Society*, vol. 6: p. 826.

— (1960*a*): "Axiom schemata of strong infinity in axiomatic set theory". *Pacific Journal of Mathematics*, vol. 10: pp. 223–238.

— (1960*b*): "Principles of reflection in axiomatic set theory". *Fundamenta Mathematicae*, vol. 49: pp. 1–10.

— (1965): "A hierarchy of formulas in set theory". *Memoirs of the American Mathematical Society*, vol. 57.

— (1979): *Basic Set Theory.* Springer-Verlag, Berlin. Reprinted by Dover Publications, Mineola, New York, 2002.

Linsky, Bernard (1999): *Russell's Metaphysical Logic*. CSLI Lecture Notes No. 101. CSLI Publications, Stanford.

Linsky, Bernard; Zalta, Edward N. (2006): "What is neologicism?" *The Bulletin of Symbolic Logic*, vol. 12: pp. 60–99.

Löwenheim, Leopold (1915): "Über Möglichkeiten im Relativkalkül". *Mathematische Annalen*, vol. 76: pp. 447–470. English translation by Stefan Bauer-Mengelberg published in van Heijenoort [1967*a*], pp. 228–251.

MacLane, Saunders (1938): "Carnap on logical syntax". *Bulletin of the American Mathematical Society*, vol. 44: pp. 171–176.

Maddy, Penelope (1988): "Believing the axioms. I". *The Journal of Symbolic Logic*, vol. 53: pp. 481–511.

— (1990): *Realism in Mathematics*. Oxford University Press, Oxford.

Malcev, Anatolii (1936): "Untersuchungen aus dem Gebiete der mathematischen Logik". *Matematicheskii Sbornik*, vol. 43: pp. 323–336.

McGee, Vann (1990): *Truth, Vagueness, and Paradox. An Essay on the Logic of Truth*. Hackett Publishing, Indianapolis.

— (1997): "How we learn mathematical language". *The Philosophical Review*, vol. 106: pp. 35–68.

Mendelson, Elliott (1958): "The axiom of Fundierung and the axiom of choice". *Archive for Mathematical Logic*, vol. 4: pp. 65–70.

Mirimanoff, Dimitry (1917): "Les antinomies de Russell et de Burali-Forti et le problème fondamental de la théorie des ensembles". *L'Enseignement mathématique*, vol. 19: pp. 37–52.

Montague, Richard; Vaught, Robert L. (1959): "Natural models of set theories". *Fundamenta Mathematicae*, vol. 47: pp. 219–242.

Moore, Gregory H. (1982): *Zermelo's Axiom of Choice. Its Origins, Development, and Influence*. Springer-Verlag, Berlin.

Mostowski, Andrzej (1947): "On definable sets of positive integers". *Fundamenta Mathematicae*, vol. 34: pp. 81–112. Reprinted in Mostowski [1979], pp. 339–370.

— (1950): "Some impredicative definitions in the axiomatic set-theory". *Fundamenta Mathematicae*, vol. 37: pp. 111–124. Reprinted in Mostowski [1979], pp. 479–492.

— (1979): *Foundational Studies. Selected Works. Vol. 1.* North-Holland Publishing, Amsterdam.

Muirhead, J. H. (ed.) (1924): *Contemporary British Philosophy: Personal Statements.* First Series. Allen & Unwin, London.

Myhill, John (1958): "Problems arising in the formalization of intensional logic". *Logique et Analyse*, vol. 1: pp. 78–83.

— (1979): "A refutation of an unjustified attack on the axiom of reducibility". In Roberts [1979], pp. 81–90.

Neale, Stephen (2001): *Facing Facts.* Clarendon Press, Oxford.

Nef, Frédéric; Vernant, Denis (eds.) (1998): *Le formalisme en question. Le tournant des années 30.* Librairie Philosophique J. Vrin, Paris.

von Neumann, John (1923): "Zur Einführung der transfiniten Zahlen". *Acta litterarum ac scientiarum Regiae Universitatis Hungaricae Francisco-Josephinae, Sectio scientarium mathematicarum*, vol. 1: pp. 199–208. English translation by Jean van Heijenoort published in van Heijenoort [1967*a*], pp. 346–354.

— (1925): "Eine Axiomatisierung der Mengenlehre". *Journal für die reine und angewandte Mathematik*, vol. 154: pp. 219–240. English translation by Stefan Bauer-Mengelberg and Dagfinn Føllesdal published in van Heijenoort [1967*a*], pp. 393–413.

— (1929): "Über eine Widerspruchfreiheitsfrage in der axiomatischen Mengenlehre". *Journal für die reine und angewandte Mathematik*, vol. 160: pp. 227–274.

Paris Logic Group, The (ed.) (1987): *Logic Colloquium '85.* Elsevier (North-Holland Publishing), Amsterdam.

Parsons, Charles (1974): "Sets and classes". *Noûs*, vol. 8: pp. 1–12. Reprinted in Parsons [1983], pp. 209–220.

— (1977): "What is the iterative conception of set?" In Butts and Hintikka [1977], pp. 335–367. Reprinted in Parsons [1983], pp. 268–297. Page references are to the reprint.

— (1983): *Mathematics in Philosophy. Selected Essays.* Cornell University Press, Ithaca, New York.

Patterson, Douglas (ed.) (2008): *New Essays on Tarski and Philosophy.* Oxford University Press, Oxford.

Peliš, Michal (ed.) (2008): *The Logica Yearbook 2007.* Filosofia, Praha.

Peressini, Anthony F. (1997): "Cumulative versus noncumulative ramified types". *Notre Dame Journal of Formal Logic*, vol. 38: pp. 385–397.

Poincaré, Henri (1905): "Les mathématiques et la logique. I". *Revue de Métaphysique et de Morale*, vol. 13: pp. 815–835. Reprinted with extensive deletions in Poincaré [1908], pp. 152–171.

— (1906*a*): "Les mathématiques et la logique. II". *Revue de Métaphysique et de Morale*, vol. 14: pp. 17–34. Reprinted with extensive deletions in Poincaré [1908], pp. 172–191, under the title 'Les logiques nouvelles'.

— (1906*b*): "Les mathématiques et la logique. III". *Revue de Métaphysique et de Morale*, vol. 14: pp. 294–317. Reprinted with extensive deletions in Poincaré [1908], pp. 192–214, under the title 'Les derniers efforts des Logisticiens'.

— (1908): *Science et méthode*. Flammarion, Paris.

Potter, Michael (2000): *Reason's Nearest Kin. Philosophies of Arithmetic from Kant to Carnap*. Oxford University Press, Oxford.

— (2004): *Set Theory and Its Philosophy. A Critical Introduction*. Oxford University Press, Oxford.

Procházka, Karel (2006): "Consequence and semantics in Carnap's syntax". In Kolman [2006], pp. 77–113.

— (2008): "Once again on ω-inferences and Tarski's definition of logical consequence". In Peliš [2008], pp. 142–156.

Quine, Willard Van Orman (1955): "On Frege's way out". *Mind, New Series*, vol. 64: pp. 145–159.

— (1960): "Carnap and logical truth". *Synthese*, vol. 12: pp. 350–374. Reprinted in Schilpp [1963], pp. 385–406. Page references are to the reprint.

— (1969): *Set Theory and Its Logic*. Revised Edition. Harvard University Press, Cambridge, Massachusetts.

Ramsey, Frank Plumpton (1925): "The foundations of mathematics". *Proceedings of the London Mathematical Society*, vol. 25: pp. 338–384. Reprinted in Ramsey [1990], pp. 164–224. Page references are to the reprint.

— (1990): *Philosophical Papers*. Cambridge University Press, Cambridge. Edited by David Hugh Mellor.

Ray, Greg (2005): "On the matter of essential richness". *Journal of Philosophical Logic*, vol. 34: pp. 433–457.

Rayo, Augustín; Uzquiano, Gabriel (eds.) (2006): *Absolute Generality*. Clarendon Press, Oxford.

Resnik, Michael D. (1966): "On Skolem's paradox". *The Journal of Philosophy*, vol. 63: pp. 425–438.

Ricketts, Thomas (1997): "Truth-values and courses-of-value in Frege's *Grudgesetze*". In Tait [1997], pp. 187–211.

— (2007): "Tolerance and logicism: Logical syntax and the philosophy of mathematics". In Friedman and Creath [2007], pp. 200–225.

Roberts, George W. (ed.) (1979): *Bertrand Russell Memorial Volume*. Allen & Unwin, London.

de Rouilhan, Philippe (1998): "Tarski et l'universalité de la logique. Remarques sur le *post-scriptum* au 'Wahrheitsbegriff'". In Nef and Vernant [1998], pp. 85–102.

— (2009): "Carnap on logical consequence for Languages I and II". In Wagner [2009], pp. 121–146.

Russell, Bertrand (1903): *The Principles of Mathematics*. Allen & Unwin, London. Second edition 1937.

— (1906*a*): "Les paradoxes de la logique". *Revue de Métaphysique et de Morale*, vol. 14: pp. 627–650. English version published in Russell [1973], pp. 190–214, under the title "On 'Insolubilia' and their solution by symbolic logic". Page references are to the English version.

— (1906*b*): "On some difficulties in the theory of transfinite numbers and order types". *Proceedings of the London Mathematical Society*, vol. 4: pp. 29–53. Reprinted in Russell [1973], pp. 135–164. Page references are to the reprint.

— (1908): "Mathematical logic as based on the theory of types". *American Journal of Mathematics*, vol. 30: pp. 222–262. Reprinted in Russell [1956], pp. 57–102. Page references are to the reprint.

— (1910): "La théorie des types logiques". *Revue de Métaphysique et de Morale*, vol. 18: pp. 263–301. English version reprinted in Russell [1973], pp. 215–252.

— (1924): "Logical atomism". In Muirhead [1924], pp. 356–383. Reprinted in Russell [1956], pp. 321–343. Page references are to the reprint.

— (1956): *Logic and Knowledge.* Allen & Unwin, London. Edited by Robert C. Marsh.

— (1973): *Essays in Analysis.* Allen & Unwin, London. Edited by Douglas Lackey.

Russell, Bertrand; Whitehead, Alfred North (1962): *Principia Mathematica to ∗56.* Cambridge University Press, Cambridge.

Savage, C. Wade; Anderson, C. Anthony (eds.) (1989): *Rereading Russell. Essays on Bertrand Russell's Metaphysics and Epistemology.* University of Minnesota Press, Minneapolis.

Schilpp, Paul Arthur (ed.) (1944): *The Philosophy of Bertrand Russell.* The Library of Living Philosophers Vol. V. Open Court, La Salle, Illinois.

— (1963): *The Philosophy of Rudolf Carnap.* The Library of Living Philosophers Vol. XI. Open Court, La Salle, Illinois.

Schirn, Matthias (ed.) (1998): *The Philosophy of Mathematics Today.* Clarendon Press, Oxford.

Scott, Dana S. (1974): "Axiomatizing set theory". In Jech [1974], pp. 207–214.

Shapiro, Stewart (1991): *Foundations without Foundationalism. A Case for Second-order Logic.* Oxford Logic Guides No. 17. Clarendon Press, Oxford.

Shapiro, Stewart; Wright, Crispin (2006): "All things indefinitely extensible". In Rayo and Uzquiano [2006], pp. 255–304.

Sharlow, Mark F. (2001): "Broadening the iterative conception of set". *Notre Dame Journal of Formal Logic*, vol. 42: pp. 149–170.

Shepherdson, John C. (1952): "Inner models for set theory – Part II". *Journal of Symbolic Logic*, vol. 17: pp. 225–237.

Sher, Gila; Tieszen, Richard (eds.) (2000): *Between Logic and Intuition. Essays in Honor of Charles Parsons.* Cambridge University Press, Cambridge.

Shoenfield, Joseph R. (1967): *Mathematical Logic.* Addison-Wesley, Reading, Massachusetts.

Sierpinski, Waclaw; Tarski, Alfred (1930): "Sur une propriété charactéristique des nombres inaccessibles". *Fundamenta Mathematicae*, vol. 15: pp. 292–300.

Skolem, Thoralf (1920): "Logisch-kombinatorische Untersuchungen über die Erfüllbarkeit oder Beweisbarkeit mathematischer Sätze nebst einem Theoreme über dichte Mengen". In *Videnskapsselskapets skrifter, I. Matematisk-naturvidenskabelig klasse*, number 4, pp. 252–263. English translation by Stefan Bauer-Mengelberg published in van Heijenoort [1967*a*], pp. 252–263.

— (1923): "Einige Bemerkungen zur axiomatischen Begründung der Mengenlehre". In *Matematikerkongressen i Helsingfors den 4–7 Juli 1922, Den femte skandinaviska matematikerkongressen*, pp. 217–232. Akademiska Bokhandeln, Helsinki. English translation by Stefan Bauer-Mengelberg published in van Heijenoort [1967*a*], pp. 290–301. Page references are to the translation.

— (1928): "Über die mathematische Logik". *Norsk matematisk tidsskrift*, vol. 10: pp. 125–142. English translation by Stefan Bauer-Mengelberg and Dagfinn Føllesdal published in van Heijenoort [1967*a*], pp. 508–524. Page references are to the translation.

Smith, Peter (2007): *An Introduction to Gödel's Theorems.* Cambridge University Press, Cambridge.

Smullyan, Raymond M.; Fitting, Melvin (1996): *Set Theory and the Continuum Problem.* Oxford Logic Guides No. 34. Clarendon Press, Oxford.

Sobociński, Bolesław (1949): "L'Analyse de l'antinomie russellienne par Leśniewski". *Methodos*, vol. 1: pp. 94–107, 220–228, 308–316.

Suppe, Frederick; Asquith, Peter D. (eds.) (1977): *PSA 1976: Proceedings of the Biennial Meeting of the Philosophy of Science Association, Vol. 2.* Philosophy of Science Association, East Lansing, Michigan.

Suppes, Patrick (1960): *Axiomatic Set Theory.* Van Nostrand, New York. Reprinted by Dover Publications, Mineola, New York, 1972.

Tait, William W. (ed.) (1997): *Early Analytic Philosophy. Frege, Russell, Wittgenstein.* Open Court, Chicago.

Tait, William W. (1998*a*): "Foundations of set theory". In Dales and Oliveri [1998], pp. 273–290.

— (1998*b*): "Zermelo's conception of set theory and reflection principles". In Schirn [1998], pp. 469–483.

— (2000): "Cantor's *Grundlagen* and the paradoxes of set theory". In Sher and Tieszen [2000], pp. 269–290.

— (2005*a*): "Constructing cardinals from below". In Tait [2005*b*], pp. 133–154.

— (2005*b*): *The Provenance of Pure Reason. Essays in the Philosophy of Mathematics and Its History*. Oxford University Press, Oxford.

Tarski, Alfred (1933): *Pojęcie prawdy w językach nauk dedukcyjnych*. Towarzystwo Naukowe Warszawskie, Warszawa. English translation by J. H. Woodger published in Tarski [1983], pp. 152–278. Page references are to the translation.

— (1935): "Der Wahrheitsbegriff in den formalisierten Sprachen". *Studia Philosophica*, vol. 1: pp. 261–405.

— (1936*a*): "O pojciu wynikania logicznego". *Przeglad Filozoficzny*, vol. 39: pp. 58–68. English translation by J. H. Woodger in Tarski [1983], pp. 409–420. Page references are to the translation.

— (1936*b*): "O ungruntowaniu naukowej semantyki". *Przeglad Filozoficzny*, vol. 39: pp. 50–57. English translation by J. H. Woodger in Tarski [1983], pp. 401–408. Page references are to the translation.

— (1944): "The semantic conception of truth and the foundations of semantics". *Philosophy and Phenomenological Research*, vol. 4: pp. 341–376.

— (1983): *Logic, Semantics, Metamathematics. Papers from 1923 to 1938*. Second edition, John Corcoran (ed.). Hackett Publishing Company, Indianapolis, Indiana.

Taylor, R. Gregory (1993): "Zermelo, reductionism, and the philosophy of mathematics". *Notre Dame Journal of Formal Logic*, vol. 34: pp. 539–563.

Tiles, Mary (1989): *The Philosophy of Set Theory. An Historical Introduction to Cantor's Paradise*. Basil Blackwell, Oxford.

Urquhart, Alasdair (2003): "The theory of types". In Griffin [2003], pp. 286–309.

Uzquiano, Gabriel (1999): "Models of second-order Zermelo set theory". *The Bulletin of Symbolic Logic*, vol. 5: pp. 289–302.

— (2002): "Categoricity theorems and conceptions of set". *Journal of Philosophical Logic*, vol. 31: pp. 181–196.

Wagner, Pierre (ed.) (2009): *Carnap's Logical Syntax of Language*. Palgrave Macmillan, Basingstoke, Hampshire.

Wang, Hao (1974): *From Mathematics to Philosophy*. Routledge and Kegan Paul, London.

Weston, Thomas S. (1977): "The continuum hypothesis is independent of second-order ZF". *Notre Dame Journal of Formal Logic*, vol. 18: pp. 499–503.

Wittgenstein, Ludwig (1922): *Tractatus Logico-Philosophicus*. Routledge & Kegan Paul, London. A German–English edition; English translation by C. K. Ogden.

Zalta, Edward N. (2009): "Frege's logic, theorem, and foundations for arithmetic". In *The Stanford Encyclopedia of Philosophy* (Edward N. Zalta, ed.). Summer 2009 edition. URL `http://plato.stanford.edu/archives/sum2009/entries/frege-logic/`.

Zermelo, Ernst (1904): "Beweis, daß jede Menge wohlgeordnet werden kann". *Mathematische Annalen*, vol. 59: pp. 514–516. English translation by Stefan Bauer-Mengelberg published in van Heijenoort [1967*a*], pp. 139–141.

— (1908): "Untersuchungen über die Grundlagen der Mengenlehre I". *Mathematische Annalen*, vol. 65: pp. 261–281. English translation by Stefan Bauer-Mengelberg published in van Heijenoort [1967*a*], pp. 199–215.

— (1929): "Über den Begriff der Definitheit in der Axiomatik". *Fundamenta Mathematicae*, vol. 14: pp. 339–344.

— (1930): "Über Grenzzahlen und Mengenbereiche: Neue Untersuchungen über die Grundlagen der Mengenlehre". *Fundamenta Mathematicae*, vol. 16: pp. 29–47. English translation by Michael Hallett published in Ewald [1996], pp. 1219–1233. Page references are to the translation.