Type-free Truth

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilians-Universität München

vorgelegt von
Thomas Schindler
aus München
2015

Erstgutachter: Prof. Dr. Dr. Hannes Leitgeb

Zweitgutachter: Prof. Dr. Volker Halbach

Datum der mündlichen Prüfung: 29.01.2015

Abstract

This book is a contribution to the flourishing field of formal and philosophical work on truth and the semantic paradoxes. Our aim is to present several theories of truth, to investigate some of their model-theoretic, recursion-theoretic and proof-theoretic aspects, and to evaluate their philosophical significance.

In Part I we first outline some motivations for studying formal theories of truth, fix some terminology, provide some background on Tarski's and Kripke's theories of truth, and then discuss the prospects of classical type-free truth. In Chapter 4 we discuss some minimal adequacy conditions on a satisfactory theory of truth based on the function that the truth predicate is intended to fulfil on the deflationist account. We cast doubt on the adequacy of some non-classical theories of truth and argue in favor of classical theories of truth.

Part II is devoted to grounded truth. In chapter 5 we introduce a game-theoretic semantics for Kripke's theory of truth. Strategies in these games can be interpreted as reference-graphs (or dependency-graphs) of the sentences in question. Using that framework, we give a graph-theoretic analysis of the Kripke-paradoxical sentences. In chapter 6 we provide simultaneous axiomatizations of groundedness and truth, and analyze the proof-theoretic strength of the resulting theories. These range from conservative extensions of Peano arithmetic to theories that have the full strength of the impredicative system ID_1.

Part III investigates the relationship between truth and set-theoretic comprehension. In chapter 7 we canonically associate extensions of the truth predicate with Henkin-models of second-order arithmetic. This relationship will be employed to determine the recursion-theoretic complexity of several theories of grounded truth and to show the consistency of the latter with principles of generalized induction. In chapter 8 it is shown that the sets definable over the standard model of the Tarskian hierarchy are exactly the hyperarithmetical sets. Finally, we try to apply a certain solution to the set-theoretic paradoxes to the case of truth, namely Quine's idea of stratification. This will yield classical disquotational theories that interpret full second-order arithmetic without set parameters, Z_2^- (chapter 9). We also indicate a method to recover the parameters.

An appendix provides some background on ordinal notations, recursion theory and graph theory.

Contents

Acknowledgements

Writing this thesis would have been impossible without the help and support of many people.

First, I want to thank my supervisor Hannes Leitgeb for all his support and encouragement during the writing of this thesis. His many comments and advices on my papers and talks were an invaluable help to me. His paper "What truth depends on" [55] had a great impact on me, as witnessed by Part II of my thesis.

A special thanks goes to Jeff Ketland, who supervised my work before I became a doctoral fellow at the MCMP. Before he left to Oxford, I had almost weekly meetings with him in which I could probe my ideas and would always get valuable feedback.

Volker Halbach was kind enough to read some of my papers and provide helpful comments on it. My conversations with him in Amsterdam and Chiemsee sharpened my views on some of the topics that I deal with in this thesis. Almost everything I know about axiomatic truth theories derives from my reading of his book [38]. I am very happy that he agreed to be the external examiner of this thesis.

I am deeply indebted to my friend Timo Beringer, with whom I had countless conversations about logic and philosophy in the last 5 years or so. He proof-read all of my papers and discussed all my ideas with me. He sharpened my views on many topics and was always very helpful when I had certain technical questions. My joint paper with him, "Reference-graphs, games for truth, and paradox" is one of the papers that I am most proud of.

I am obliged to Lavinia Picollo, with whom I had many discussions about metamathematics and deflationism in the last year. She certainly influenced my thinking about these topics very much. I am very proud of our joint paper "Deflationism and the purpose of truth".

Stanislav Speranski carefully read chapters 6-8 of this thesis and provided helpful suggestions. We had several meetings where we discussed the contents of chapter 7, which allowed me to improve this chapter significantly.

I thank my colleagues at the MCMP for providing such a stimulating enviroment. A special thanks goes to Catrin Campbell-Moore, Johannes Stern and Martin Fischer with whom I had several discussions about the topics in this thesis. A thanks goes also out to my office colleagues, Johannes Korbmacher, Hans-Christoph Kotzsch, Albert Anglberger and Thomas Meier.

I thank Eduardo Barrio and the logic group of Buenos Aires, who made my stay in Buenos Aires a real pleasure. I thank the DAAD for supporting that trip financially.

Thanks to Jonas Raab and Christian Meyer for listening to all of my stories, and thanks to Jonas for proof-reading so many of my papers.

I also want to take the opportunity to thank my teachers in logic, Godehard Link, Marek Polanski and Karl-Georg Niebergall.

I have presented my work at several workshops and conferences, including Vienna, Chongching, Amsterdam, Bristol, Oslo, Canterbury, Chiemsee, Buenos Aires and, of course, Munich. I thank the attendees of these conferences for their feedback. In particular, I thank Leon Horsten, Graham Leigh, Toby Meadows, Julien Murzi, David Ripley, Jönne Speck, and Philip Welch.

I am very grateful to Stephan Hartmann and Dieter Donder for having agreed to be on the committee of my defense.

This work was generously supported by the Alexander von Humboldt foundation and the Munich Center for Mathematical Philosophy. I could not have written this thesis without their financial support.

Last but certainly not least I want to thank my family, Mama, Armin, Tanja, Tina and Toni for all their support during the last years.

Co- and single-authored publications

Sections 4.1-4.2 and parts of section 1.1 were jointly written with Lavinia Picollo from the University of Buenos Aires. The material is part of our paper [65], which is currently in preparation. Of course, I take full responsibility for the way the material is presented here.

Sections 5.3-5.4.3 were jointly written with Timo Beringer from the LMU Munich. The material is part of our paper [7], which is currently in preparation. The paper contains a lot of additional material that has not been included in this thesis. Of course, I take full responsibility for the way the material is presented here.

All other parts of this thesis were written solely by myself. All proofs in this book were carried out solely by myself, except those that resulted from joint work with Ms. Picollo or Mr. Beringer, as indicated above. A few theorems that I have proved are implicit in the work of others, and I have indicated the relevant paper in such cases. All theorems in this book that were not proved by myself are stated without proofs, and I have indicated the owner of the result in brackets.

Some of the results in sections 6.2-6.3 have been previously published in my article Schindler [81]. I have used the opportunity to correct some flaws in the paper and added two further axiomatic systems, called WKG and MG. Proofs were mostly omitted from the paper; they are given here for the first time. Overall, my views have changed quite a bit since the writing of the article and I deem the present exposition superior in many respects.

Sections 9.2-9.3 are based on my publication Schindler [82]. In the paper, I did not have enough space to properly motivate the systems introduced there. In the present book I fill this gap.

Part I.

Truth as a logico-mathematical notion

1. Introduction

In the nineteen-thirties, Tarski [89] showed how to give a rigorous definition of truth for a range of interpreted formal languages. His truth definition involves two languages: first, the language for which truth is defined (the object language), and second, the language in which the definition is given (the metalanguage). The latter must be 'rich' enough to talk about the expressions of the object language and syntactic operations on them. Tarski took it as a minimal adequacy condition on a satisfactory definition of truth that it implies all instances of the so-called T-schema

$$T\ulcorner\varphi\urcorner \leftrightarrow \varphi,$$

where $\ulcorner\varphi\urcorner$ is a name of the sentence φ. Under fairly minimal conditions, object- and metalanguage cannot coincide, on pain of contradiction. This is known as Tarski's undefinability theorem, which can be proved by formalizing the infamous liar paradox. Consider the following sentence:

$$\text{The sentence marked (1) is not true} \tag{1}$$

The assumption that (1) is true leads to the conclusion that (1) is not true and vice versa. The Tarskian truth predicates are *typed* truth predicates, in the sense that they provably apply only to sentences of the object language. If we want a truth predicate for the metalanguage, we have to move up one level to a meta-metalanguage, and so forth, thus creating the Tarskian hierarchy of languages and truth predicates.

Tarski's work was a huge success and paved the way for model theory. However, not everyone was content with Tarski's semantics. Philosophers strive for the absolute; if we say 'All sentences of the form 'If p then p' are true', we want the quantifier to range over *all* sentences, even those that contain an occurrence of the truth predicate. Since the mid-seventies, philosophical logicians have increasingly tried to close the gap between object- and metalanguage, resulting in a variety of semantic and axiomatic theories of *type-free* truth. Work in this area includes Kripke [53], Friedman & Sheard [28], Feferman [19] [20] [21], Cantini [10] [11], Herzberger [42], Gupta [31], and Belnap & Gupta [32].

On the semantic approach, one usually starts with a model for the language without the truth predicate and then assigns an interpretation (extension) for the truth predicate such that certain plausible principles of truth are satisfied (e.g.

instances of the T-schema). On the axiomatic approach, such principles are studied directly from a proof-theoretic point of view. Investigations in this area show that the addition of a truth predicate to a language increases the expressive power of that language in several ways. On the semantic side, we observe that new sets can be defined, or that quantification over subsets of the domain becomes feasible. On the axiomatic side, the truth predicate enables us to finitely axiomatize infinite sets of sentences, shorten the proofs of theorems, or prove new theorems.

In this increase in expressive power, proponents of the deflationary account of truth see the sole reason why we have a truth predicate in our language at all. According to that view, once we have understood the function of the truth predicate, we understand about everything there is to know about truth. No definition of the form 'x is true if and only if p' is required. Deflationism is currently the most popular philosophical account of truth, being "attractively demystifying" (Horwich [46, p. 5]). Although I do not fully endorse deflationism, I agree with Field that we should be *methodological* deflationists:

> [W]e should start out assuming deflationism as a working hypothesis; we should adhere to it unless and until we find ourselves reconstructing what amounts to the inflationist's relation '[Sentence] S has truth conditions p'. (Field [22, p. 367])

Deflationists usually emphasize the role of the truth predicate for expressing generalizations or 'infinite conjunctions', but philosophers have assigned other purposes to the truth predicate too, for example, in the philosophy of mathematics. Russell [80] observed that the notion of truth provides us with virtual classes, and thus can be used for ontological reductions. Loosely speaking, the claim that x is an element of the set $\{y|\varphi(y)\}$ is intertranslatable (model-theoretically and proof-theoretically) with the claim that $\varphi(x)$ is true. This is the reason why the addition of a truth predicate allows us to mimick second-order quantification.

The aim of this essay is to explore if, how and to what extent the truth predicate can serve the purposes that philosophers have assigned to it in face of the semantic paradoxes. Let us therefore have a closer look at some of the tasks that we want the truth predicate to perform.

1.1. Deflationism

Deflationism seems to have originated with the writings of Frege [27], Ramsey [76] and Quine [72]; its modern champions include Horwich [46] and Field [22], among many others. Deflationists claim that the truth predicate exists *solely* for the sake of certain logico-linguistic or logico-mathematical purposes and would otherwise be entirely dispensable. According to that view, the truth predicate provides a means of capturing a (possibly infinite) set of sentences by a single expression without exhibiting tokens of the original sentences. We can construct a single new statement, closely related to the original sentences, either by applying the truth predicate to definite descriptions or proper names of the original sentences or by subsuming a predicate that applies exactly to these sentences under the truth predicate.

For example, instead of using the sentence 'The universal proportionality factor between equivalent amounts of energy and mass is equal to the speed of light squared' we can choose a definite description of that sentence—such as 'the most famous formula of physics'—or a proper name—such as 'Einstein's mass-energy equivalence'—and use the sentence 'Einstein's mass-energy equivalence is true' instead. Definite descriptions and proper names denote just one object. They enable us to formulate what we may call *singular truth ascriptions*, expressions of the form 's is true', where s denotes a sentence, without displaying that sentence.[1]

On the other hand, by supplying a property that all sentences in a certain set share, we can capture all those sentences in a single phrase by formulating what we may call a *general truth ascription*, or *generalization* for short. These are expressions of the form 'All Φs are true'. For instance, instead of repeating all of Newton's three laws of motion one by one, we can simply say 'All of Newton's law of motion are true'.

Of course, singular truth ascriptions can be seen as a limit case of generalizations, where the anteceding property is satisfied just by one sentence. It is always possible to replace sentences of the form 's is true' by their logical equivalents 'All sentences identical to s are true'. As a consequence, in what follows we will focus on general truth ascriptions.

Capturing an infinite set of sentences by a single expression can be seen as a version of finite axiomatizability. The insight that the truth predicate can be used for finite axiomatizations is actually not a discovery of deflationists. It is well-known that important mathematical theories like Peano arithmetic or Zermelo-Fraenkel set theory are not finitely axiomatizable. Kleene [50] and later Craig & Vaught [16] showed that almost any theory is finitely axiomatizable if additional predicate symbols are allowed in the axiomatization. That is, if $\mathcal{S}$ is a theory with finite non-logical vocabulary that has infinite models only, there is a conservative extension of $\mathcal{S}$ that is finitely axiomatizable. Roughly, the strategy here is to introduce a truth (or satisfaction) predicate governed by the Tarski-clauses and then add the statement 'All axioms of $\mathcal{S}$ are true'.

The *reasons* why it is convenient to have such an expressive device at our disposal are well known. We can roughly divide them into three categories—epistemic, rhetoric and logical—of which the last one is without question the most important

[1] Of course, the truth predicate can also be applied to the quote-name of a sentence, but this case is rather uninteresting. For when using them we exhibit a token of the original sentence. As we will see in what follows, the truth predicate becomes handy when we don't want to or simply cannot exhibit such tokens.

one. First, we might want to adopt a certain attitude towards a set of sentences without knowing which sentences exactly belong to that set. For example, suppose yesterday you had a conversation with an expert in physics who convinced you of something, but the matter was so complicated that you don't remember exactly what she said. Then the truth predicate allows you to express your agreement with the expert by saying 'Everything the expert said yesterday is true'. The truth predicate allows us to make *blind ascriptions.*

Second, we might want to save time or space. For example, instead of repeating all the claims made in the bible, which would take an awful lot of time, we just say 'Everything the bible says is true'.

Finally, and most importantly, there are cases where it is simply impossible for us to explicitly state all the sentences in a certain set, namely when the set is infinite. We might want to affirm all theorems of Peano arithmetic, all propositional tautologies, etc. In those cases where the sentences in question are definable by a formula, we can assert them all at once by saying, e.g. 'All theorems of arithmetic are true'. Thus, Quine famously said:

> We may affirm the single sentence by just uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences then the truth predicate has its use. ([72, p. 12])

Expressing generalizations can be useful for many purposes. Obviously, they are helpful in stating the laws of logic (as in the above example), but they also enable us, amongst other things, to express agreement[2] with theories that cannot be finitely axiomatized (except by using the truth predicate, that is) or to make commitments explicit. For example, it is generally held that someone who believes all theorems of Peano arithmetic (PA) should also believe that PA is sound. Since PA does not contain its own truth predicate, this commitment is usually expressed by the schematic *local* reflection principle

$$Prov_{PA}(\ulcorner\varphi\urcorner) \to \varphi$$

where $Prov_{PA}(x)$ is the standard provability predicate of PA. By adding a truth predicate T to the language, we can express all the instances of the local reflection principle in a single sentence, namely, a formalized version of 'All theorems of PA are true', i.e.

$$\forall x(Prov_{PA}(x) \to Tx)$$

[2] Here, agreement is understood to be more than just the autobiographical assertion that e.g. someone believes all the theorems of PA—the autobiographical claim 'I believe all theorems of PA' might be true while some theorems of PA are false. As Field [26] puts it, expressing agreement with PA is making a claim that is correct if and only if PA is correct.

The latter is called *global reflection principle* for PA. It implies, under minimal conditions, the consistency of PA.

Another important purpose of generalizations is that they allow us to express disagreement with non-finitely axiomatizable theories. We often disagree with some theory without knowing exactly where it goes wrong. If the theory in question is finitely axiomatized, we can express our disagreement by disjoining the negations of the individual axioms of the theory. But in the case of a theory with infinitely many axioms the only way to express our disagreement is by saying: 'Not everything in the theory is true' or 'Something in the theory is false'.

At this point, it is important to remark, however, that the truth predicate can serve the purpose of expressing agreement and disagreement only to a certain extent—*even* if the truth predicate is fully transparent (i.e. if a sentence and its truth predication are intersubstitutable *salva veritate* in every transparent context). For there are cases in which we can't express our argreement or disagreement *consistently.* Suppose, for example, that Jones says 'The Liar sentence is not true' (=the Liar), and assume furthermore that Brown agrees with Jones. Now it seems that Brown can express her agreement with Jones by saying 'What Jones said is true', because (assuming transparency) the latter will be materially equivalent to what Jones said. However, by the way that the Liar sentence is defined, Brown's utterance is *also* equivalent to the negation of what Jones said—Brown's utterance is equivalent *both* to the Liar and its negation.

One should not underestimate the importance of this example. For one argument that has often been raised against classical truth theories (i.e. truth theories based on classical first-order logic) is that they cannot accommodate the unrestricted T-schema and therefore (so the argument goes) compromise the role of truth (e.g. Field [26]). The above example shows, I hope, that the liar sentence will place certain restrictions on *any* theory of truth, regardless of the background logic.

The discussion so far leaves open the question which principles the truth predicate has to validate in order to fulfill the generalizing function. Adding truth as a primitive predicate symbol to our language certainly allows us to syntactically *formulate* expressions such as $\forall x(\varphi(x) \to Tx)$, but this is completely useless if the truth predicate is not governed by axioms that relate a generalization in an appropriate way to the sentences that we want to capture with it.

The most popular view on deflationist theories of truth is *disquotationalism*, i.e. the idea that all there is to say about truth is exhausted by the equivalence between φ and $T\ulcorner\varphi\urcorner$ for every sentence φ (and therefore, that the equivalence accounts for all uses of the truth predicate, in particular for its generalizing function). If the equivalence is expressed in the object language, we get the celebrated T-schema,

$$T\ulcorner\varphi\urcorner \leftrightarrow \varphi$$

Its rule-form variant is called the *Intersubstitutivity Principle*, according to which

φ and $T\ulcorner\varphi\urcorner$ entail each other.[3] While in classical logic these two principles are equivalent, this is not the case for every non-classical logic. Some non-classical theories of truth satisfy one of them but not the other. In logics with conditional proof or the rule of introduction of the conditional, the Intersubstitutivity Principle implies the T-schema, while if Modus Ponens holds the latter entails the former. Systems in which the Intersubstitutivity Principle holds are usually called theories of *transparent* truth.

For most disquotationalists, the equivalence between φ and $T\ulcorner\varphi\urcorner$ has a more-or-less analytic status. According to Field [22], the statement that φ is true is cognitively equivalent to the statement φ itself. For Horwich [46], the instances of the T-schema jointly exhaust or fix the meaning of the concept of truth.[4]

Most deflationists explicitly reject type-restrictions (cf. Horwich [46, p. 41], [47, p. 81]). However, early discussions of deflationism proceeded mostly against the background of typed theories of truth, usually an axiomatic system based on the restricted Tarski-biconditionals; only in recent years attention has gradually shifted to stronger, untyped theories of truth (cf. Halbach & Horsten [39], Halbach [38], Horsten [45]). The main reason for the initial focus on typed theories seems to be that philosophers wanted to avoid the intricacies posed by the liar paradox. Though deflationists realized that the liar will force certain restrictions and exceptions, the matter was not taken very seriously.

> Because of the paradoxes, exceptions must be made for some utterances u that contain 'true'; I won't be concerned here with just how the exceptions are to be carved out. (Field [22, p. 353, fn 1])

> There is no reason to suppose that the minimalist answers that are advanced in this essay could be undermined by any particular constructive solution to the paradoxes—so we can temporarily set those problems aside. (Horwich [46, p. 42])

In recent years philosophers have come to acknowledge that the liar paradox might pose a bigger threat to the deflationist (and in particular, the disquotationalist) account than initially thought (as witnessed e.g. by the collection Beall & Armour-Garb [2], which is wholly devoted to that problem). Disquotationalism is somewhat

[3] Field [26, p. 12] gives a more complicated formulation of the Intersubstitutivity Principle according to which if two sentences are such that the former is the result of replacing all the occurrences of the subsentence φ in transparent contexts with $T\ulcorner\varphi\urcorner$ in the latter, then both sentences entail each other. For simplicity reasons we will stick to our formulation. In most cases it is enough and implies this more complex version.

[4] Field's theory is also known as 'pure disquotationalism', while Horwich's position is usually referred to as 'minimalism'. Horwich actually does not talk about the T-schema but about the *equivalence schema* $\langle p\rangle$ is true iff p, where p ranges of propositions rather than sentences. This difference won't play a big role in our discussion.

in tension with our desire to have a truth theory based on classical logic. As Tarski's undefinability theorem shows, no classical theory that can talk about its own syntax can accommodate all instances of the T-schema. The desire to keep the unrestricted T-schema led quite a few philosophers and logicians to propose theories of truth based on some non-classical logic (e.g. Field [24], [25], [26], Beall [4], Priest [68], Weir [91], Cobreros et al. [13]). On the other hand, disquotationalists like Horwich who want to keep to classical logic have come under fire. One of the principal aims of this thesis is to investigate if and to which extent disquotationalism (and deflationism in general) is compatible with classical logic.

1.2. Virtual classes

Quine [74] famously argued that in accepting a theory, we accept its ontology; we are committed to the existence of the objects postulated by the existential statements of the theory. The truth predicate allows us to engage in class talk *without* thereby committing us to the existence of classes. An early example can be found in the work of Bertrand Russell, long before the first deflationist accounts have been formulated. After discovering the set-theoretic paradox that now bears his name, Russell tried to find a new foundation for mathematics. One radical solution, called the 'no-classes theory', was to dispense with classes altogether.

According to the no-classes theory, talk about classes has to be viewed as a *façon de parler*. Any statement involving classes must be rephrased in a way that does not explicitly mention classes. For example, the statement that Socrates is a member of the class of human beings might be expressed by saying that Socrates is human. The statement that the class of humans is not empty might be rephrased as 'There are human beings'. Working along these lines one can develop the ordinary concepts of Boolean class algebra (subset, intersection, union, complement etc.) and derive the laws that govern them. But as Quine [71] observes, one does not get much further than this, since quantification over classes cannot be mimicked in this way.

A more promising approach was explored by Russell [80] in his *substitutional theory of classes and relations* from 1906,[5] where classes are treated as incomplete and non-denoting symbols, a method that has its roots in Russell's analysis of definite descriptions [78]. On the latter, the proposition 'The present king of France is bald' is not analyzed into 'the present king of France' (subject) and 'is bald' (predicate) but rather into 'There is exactly one man who is king of France and that man is bald'. Assuming that the phrase 'the present king of France' has an independent meaning leads to a "false abstraction". Russell's proposal is to treat expressions like 'the number 1', 'the class of wise men', and 'the continuum' as false abstractions too.

[5] For a more thorough description of this theory, see Landini [54].

Russell took the quaternary relation 'q results from substituting b for a in p' as primitive (where p, a, b, q are variables that can be bound by a quantifier)—in symbols $p/a; b!q$. This might be written in a more transparent way as $p(b/a) = q$. Furthermore, let us write $p(b/a)$ for the unique q such that $p(b/a) = q$. On the intended reading, the variables range over propositions and individuals. Now, the proposition *Plato is wise* is the result of substituting Plato for Socrates in *Socrates is wise*.[6] Hence, that Plato is a member of the class of wise beings can be expressed as follows:

> The result of substituting Plato for Socrates in *Socrates is wise* is true.

The above expression is a special case of what we called a singular truth ascription. The phrase 'The result of substituting Plato for Socrates in *Socrates is wise*' denotes a proposition; but we need the truth predicate to assert it.

Russell realized that the class $\{x|x \text{ is wise}\}$ can be represented by the pair *Socrates is wise*/Socrates (which Russell calls a 'matrix'). More generally, we define

$$x \in p/a \text{ iff } p(x/a) \text{ is true}$$

and

$$p/a = q/b \text{ iff } \forall x(x \in p/a \leftrightarrow x \in q/b)$$

The matrix or 'class' p/a is an incomplete symbol, governed by contextual definitions. A proposition mentioning a matrix (class) is only significant if it can be rephrased in the basic language, that is if it can be transformed into a statement that does not mention any matrices at all. By using iterated substitutions, we can also represent relations of higher arity. For example, the binary relation $\{(x, y)|x \text{ is the father of } y\}$ can be represented by the matrix *Philipp is the father of Alexander*/Philipp, Alexander. Russell calls a matrix of the form p/a a matrix of the first type, $p/a, b$ a matrix of the second type, $p/a, b, c$ a matrix of the third type etc. Membership between classes (of the first and second type) can now be defined by setting $p/a \in q/b, c$ iff $q(p/b, a/c)$ is true, and accordingly for membership between classes of higher types. Russell's definition of elementhood creates in effect a *simple hierarchy of types.* If α is a matrix of type i and β a matrix of type j, then the expression $\alpha \in \beta$ is significant if and only if $j = i + 1$. This blocks Russell's paradox: the expression $p/a \notin p/a$ is not significant, as it cannot be reformulated in the base language.

The usual concepts of set theory can now be developed in a straightforward way. Russell defines the cardinal 0 as the class of all classes (of the first type) that do

[6] According to Russell, a proposition like *Socrates is wise* contains the object Socrates itself rather than the name or concept *Socrates*.

not contain an element, the cardinal 1 as the class of all singletons (of the first type), 2 as the class of all pairs (of first type) etc. More precisely, choose p and a such that for all x, $p(x/a)$ is false. For example, let p be the proposition that Socrates is not identical with Socrates, and let a be Socrates. Then the number 0 may be defined as the matrix $\{\forall x(p(x/a) \text{ is false})\}/p, a$.[7] Then for all q, r we find that the class q/r belongs to 0 if and only if q/r has no members. For the number 1, choose some matrix p/a that has exactly one member. Then we may set $1 := \{\exists y \forall x(p(x/a) \text{ is true} \leftrightarrow x = y)\}/p, a$. One easily verifies that the class 1 contains exactly those classes (of the first type) that have exactly one element.

Russell showed how to reduce existence assumption about sets and propositional functions to propositions and certain operations on them such as substitution. Instead of working with propositions, one can simply work with sentences and open formulas. For example, suppose that we have a name $\ulcorner\sigma\urcorner$ for every expression σ of our language and function symbols corresponding to certain syntactic operations on them. In particular, assume that we have a function symbol s such that $s(\ulcorner\varphi(a)\urcorner, \ulcorner b\urcorner) = \ulcorner\varphi(b/a)\urcorner$. Finally, assume that there is a function (symbol) n that maps every object x to some standard name $n(x)$. Then, given the *uniform* T-biconditional

$$\forall x(Ts(\ulcorner\varphi\urcorner, n(x)) \leftrightarrow \varphi(x))$$

the syntactic object $\ulcorner\varphi\urcorner$ can play the role of the class $\{x|\varphi(x)\}$—provided we have the uniform T-biconditional for the formula φ at our disposal.

And here, of course, is where the trouble comes from. A theory that can express its own syntax (or some relevant part of it) is able to formulate sentences that assert their own untruth. The T-biconditionals for such sentences render the system inconsistent in classical logic. This is actually what happened to Russell's substitutional theory of classes and relations. Although it provides a 'solution' to the set-theoretic paradoxes (in the sense that e.g. Russell's paradox cannot be formulated in the system), the theory is still inconsistent. The presence of the substitution function together with a truth predicate and a term-forming operator renders Russell's theory subject to a liar-like paradox. One of the things that we will be interested in in this book is the question how much set theory can be developed in a theory of truth that contains only some instances of the T-schema. As we will see, questions about the expressive power of truth are closely connected to the question about how much set theory is encoded in the truth predicate.

Notice that a transparent theory of truth, i.e. a theory based on some non-classical logic that accommodates all instances of the uniform T-schema in a non-trivial way, will be able to derive (interpret) *all* axioms of type theory, by the method sketched above. However, this 'achievement' is massively diminished by the fact that we cannot reason classically with these axioms. Although we have all axioms of type

[7] Here, the brackets are a term-forming device that turn a well-formed formula into a term.

theory available, we cannot even interpret very weak subsystems of second-order arithmetic in these theories. This shows that the expressive power of a truth theory does not depend solely on the truth-theoretic axioms but also on the underlying logic.

1.3. Overview of the thesis

In chapter 2 we fix some preliminary technical matters. In order to study the notion of truth we need a language that contains names for its expressions and function symbols for certain operations on these expressions. We follow the tradition and use the language of Peano arithmetic for that purpose. We assume that the reader is familiar with that system and use section 2.1 only to fix some terminology. It is convenient to assume that PA contains function symbols for certain primitive recursive functions as primitives. The choice of the base language really matters. In section 2.2 we will show that certain truth-theoretic axioms are inconsistent over PA when the language contains symbols for some primitive recursive functions, while they are consistent over PA when the language does not contain these function symbols among its vocabulary.

As we have seen, any theory of truth has to deal in one way or the other with the liar and its kind. In chapter 3, we briefly discuss (and argue against) two standard ways of evading the semantic paradoxes, namely typing (Tarski) and weakening classical logic (Kripke). Later on, Tarski's theory will be used to measure the proof-theoretic strength of type-free theories while Kripke's theory will form the starting point for our analysis of the semantic paradoxes. The latter will also serve as an inspiration for certain axiomatic theories of truth that we will introduce later. Again, we assume that the reader is familiar with most of the material and use this chapter only to fix some terminology for later reference. Observation 3.1.2 (which strengthens an important result by Halbach) should be new, though.

Chapter 4 deals with some problems that the liar poses for classical type-free truth. We have seen that the main function of the truth predicate is to enable us to express infinite conjunctions. Several authors claim that the truth predicate can serve its expressive function only if it is fully disquotational—i.e. it satisfies the general equivalence between a sentence and its truth predication, which is impossible in classical logic. We put forward a concise formulation of what it takes for a theory of truth to enable us to express infinite conjunctions and examine existing truth theories in this light. We conclude (i) that there is no need to adopt a non-classical logic—in fact, some non-classical theories of truth are clearly inadequate—and (ii) that any reasonable classical truth theory should contain T-Out among its principles. However, Hartry Field [26, chap. 7] has argued that T-Out theories have problems with expressing agreement and disagreement. In particular, T-Out theories are inconsistent with their own global reflection principle (i.e. the statement that all theorems of $\mathcal{S}$ are true), which is usually taken to express the soundness of a theory. We argue that these problems can be overcome by adopting a revised version of the global reflection principle, namely the statement that no theorem of $\mathcal{S}$ is false (section 4.3).

Part II is largely devoted to the semantic paradoxes and grounded truth.

In chapter 5 we introduce a game-theoretic semantics for Kripke's theory of truth. Strategies in these games can be interpreted as reference-graphs of the sentences in question. Using that framework, we give a graph-theoretic analysis of the Kripke-paradoxical sentences. Our proposal is to identify the set of sentences that a sentence refers to with its dependence set in the sense of Leitgeb [55]. In section 5.2 we first introduce the basic concepts of Leitgeb's paper on semantic dependence and then show that Leitgeb's theory can be treated within the framework of Kripke's fixed-point semantics. In section 5.3, we show how to define unique reference-graphs (called 'sensitivity-graphs') for those sentences that do possess a canonical dependence set and prove some theorems concerning (ω-)consistent subsets of the T-schema in terms of sensitivity. In section 5.4.1 we define a grounding game $\mathcal{G}_G(\varphi, S)$ such that φ is grounded in S if and only if player ($\exists$) has a winning strategy in the game $\mathcal{G}_G(\varphi, S)$. We then show how the strategies available in this game can be used to define an infinite family of reference-graphs for the sentence in question. These reference-graphs can be seen as a generalization of the sensitivity-graphs of section 5.3. We then use our machinery to show that a sentence is grounded if and only if it has a well-founded reference-graph. In section 5.4.2 we define a verification (falsification) game such that φ is true (false) in the fixed-point generated by the partial model $\mathcal{F}$ if and only if player ($\exists$) has a winning strategy in the verification (falsification) game for φ and $\mathcal{F}$. In section 5.4.3 we apply our machinery to obtain some graph-theoretic descriptions of the Kripke-paradoxical sentences. We show, amongst others, that if a sentence is Kripke-paradoxical, then each of its reference-graphs contains either a directed cycle or infinitely many so-called double paths.

In chapter 6 we will search for axiomatizations of the Kripkean fixed points. In addition to the truth predicate, we will introduce a new primitive predicate symbol G, intended to express 'x is grounded', and provide simultaneous axiomatizations of groundedness and truth for several Kripkean fixed points. We will provide a list of grounding axioms that mirror the inductive process by which the fixed points are generated, plus the T-schema and the compositional axioms for T restricted to G. The idea is that instead of choosing between equally plausible but jointly inconsistent truth axioms, we adopt all of them, but restrict them in a uniform manner. We also introduce a disquotational theory of grounded truth that is inspired by an article of Horwich. The main part of this section is the analysis of the proof-theoretic strength of these theories. We will show that the axiom systems for the Weak Kleene, Strong

Kleene and Leitgeb valuation scheme are able to define the truth predicates of the Tarskian hierarchy up to (but excluding) level ϵ_0, while the axiom system for the supervalautional scheme has the full strength of the impredicative theory ID_1. The system based on Horwich's notion of grounding is conjectured to be conservative over Peano arithmetic, but I have no proof of this.

We have already seen that the truth predicate allows us to code up sets by formulae, using the translation of Russell discussed in section 1.2. In Part III we investigate this relationship in a more systematic manner, both from a model-theoretic and a proof-theoretic point of view.

In chapter 7 we will first show how to canonically associate, with any extension (interpretation) of the truth predicate (which we call a 'truth-set'), a structure (interpretation) for the language of second-order arithmetic. Second, we will give a translation of the language of second-order arithmetic into the language of truth. We will show that the translation of a second-order sentence is true relative to a truth-set if and only if the original sentence is true relative to the second-order structure associated with the truth-set. This correspondence can be used for quite a few interesting recursion-theoretic and proof-theoretic purposes. We will show that if $S = (S^+, S^-)$ is the *minimal* Kripke fixed point under an appropriate valuation scheme, then S^+ is able to define fixed points of positive operators. This implies that S^+ is Π^1_1-hard and that $(\mathbb{N}, S^+)$ is a model of (the translation of) the theory ID_1. For the minimal fixed points under the Strong Kleene and the supervaluational scheme, these results have already been shown by Cantini (cf. [10], [11]). The main innovation here is that our proof also applies to Leitgeb's theory of truth. In addition, we relate the minimal Kripke fixed points to the collection of hyperarithmetical sets. Finally, we prove some interesting theorems about the theory of positive disquotation.

In chapter 8 we will show, using techniques from the previous chapter, that the sets definable over the standard model of the Tarskian hierarchy are precisely the hyperarithmetic sets. This result has been established previously by Halbach [33]. We give a slightly different proof based on the methods of the previous chapter.

In chapter 9 we utilize the translation to establish the consistency of disquotational theories of truth that are obtained by translating comprehension axioms into T-biconditionals. These results show that disquotational theories of truth can be much stronger than our best compositional theories of truth. In particular, we present a disquotational theory of truth that interprets full second-order arithmetic, Z_2^-. The minus indicates that free set parameters are not allowed in the comprehension axioms. Finally, we indicate a method to recover the parameters. In an appendix we provide some background on ordinal notations, recursion theory and graph theory.

2. Technical preliminaries

In order to study the notion of truth, and to put it to use, we need a language that contains names for its expressions and function symbols for certain operations on these expressions. We follow the tradition and use the language of Peano arithmetic for that purpose. We assume that the reader is familiar with that system and use the next section only to fix some terminology. It is convenient to assume that PA contains function symbols for certain primitive recursive functions among its vocabulary. Some interesting effects of that decision are illustrated in the second section of this chapter.

2.1. Peano arithmetic

The language of Peano arithmetic, $\mathcal{L}_{PA}$, is a first-order language that contains a denumerably infinite set of individual variables $v_0, v_1, v_2, \ldots$, the connectives $\neg, \vee$ and $\wedge$, the quantifiers $\forall$ and $\exists$ and the identity symbol $=$. We assume that all other connectives are defined in the usual way. The sole non-logical symbols are the individual constant $\overline{0}$, the unary function symbol S for the successor function, the binary function symbols $+$ and $\cdot$ for addition and multiplication, respectively, and function symbols for certain primitive recursive (p.r.) functions that we are going to specify in the course of the book. If h is such a p.r. function, we write $\underset{\cdot}{h}$ for the corresponding function symbol. The language $\mathcal{L}_T$ is obtained from $\mathcal{L}_{PA}$ by augmenting the latter with the unary predicate symbol T.

The theory PA contains the defining axioms for zero, successor, addition, multiplication and the other p.r. function symbols together with all instances of the induction axiom scheme

$$\varphi(\overline{0}) \wedge \forall x(\varphi(x) \to \varphi(Sx)) \to \forall x\varphi(x)$$

where $\varphi(x)$ is a formula of $\mathcal{L}_{PA}$. The theory PAT is obtained from PA by extending the induction axiom scheme to the full language $\mathcal{L}_T$. Notice that PAT is a conservative extension of PA.

If n is a number, we write $\overline{n}$ for its numeral, i.e. the term that is obtained by applying the symbol S n-many times to the constant $\overline{0}$. We assume some natural (standard) Gödelcoding of the expressions of $\mathcal{L}_T$. If σ is some expression, we write $\#\sigma$ for its code and $\ulcorner\sigma\urcorner$ for the numeral of its code. We occasionally identify expressions with their codes.

The formulation of the so-called uniform T-schema involves some subtleties, to which we now turn. Let $s^k_i(m,n) = \#\varphi(\overline{n}/x_j)$, provided that $m = \#\varphi$ is a formula with exactly k free variables and x_j is its i-th free variable (according to the index ordering). The functions s^k_i are primitive recursive and will be represented by the symbols $\underset{.}{s}^k_i$ (with a subdot). Given $\varphi := \varphi(x,y,z)$ with exactly x,y,z free and index$(x) <$ index$(y) <$ index(z), we write $\ulcorner\varphi(\dot{x},\dot{y},\dot{z})\urcorner$ for $\underset{.}{s}^1_1(\underset{.}{s}^2_2(\underset{.}{s}^3_3(\ulcorner\varphi\urcorner,z),y),x)$, and similarily for formulae with n free variables. We often write $\underset{.}{s}$ instead of $\underset{.}{s}^1_1$. Then the uniform T-schema can be written as

$$\forall x_1 \ldots \forall x_n (T\ulcorner\varphi(\dot{x_1} \ldots \dot{x_n})\urcorner \leftrightarrow \varphi(x_1,\ldots,x_n))$$

Furthermore, we assume that $\mathcal{L}_{PA}$ contains the unary function symbols $\underset{.}{\neg}$ and $\underset{.}{T}$ and the binary function symbols $\underset{.}{=}, \underset{.}{\wedge}, \underset{.}{\vee}, \underset{.}{\forall}$ such that the following is derivable for all terms s,t and formulae φ,ψ:

$$\vdash \ulcorner s\urcorner \underset{.}{=} \ulcorner t\urcorner = \ulcorner s = t\urcorner$$

$$\vdash \underset{.}{\neg}\ulcorner\varphi\urcorner = \ulcorner\neg\varphi\urcorner$$

$$\vdash \ulcorner\varphi\urcorner\underset{.}{\wedge}\ulcorner\psi\urcorner = \ulcorner\varphi\wedge\psi\urcorner$$

$$\vdash \ulcorner\varphi\urcorner\underset{.}{\vee}\ulcorner\psi\urcorner = \ulcorner\varphi\vee\psi\urcorner$$

$$\vdash \underset{.}{\forall}(\ulcorner v_i\urcorner,\ulcorner\varphi\urcorner) = \ulcorner\forall v_i\varphi\urcorner$$

$$\vdash \underset{.}{T}t = \ulcorner Tt\urcorner$$

The evaluation function *val* that applied to (the code of) a closed term t gives the value (denotation) of t is primitive recursive and will be represented by the formula $y^\circ = x$.

We let $Sent_T(x)$ naturally represent the set of (codes of) $\mathcal{L}_T$-sentences, $Fm_T(x)$ the set of $\mathcal{L}_T$-formulae, $ClTerm(x)$ the set of closed terms and $Var(x)$ the set of variables. We let $Sent_{PA}(x)$ represent the set of $\mathcal{L}_{PA}$-sentences and $Fm_{PA}(x)$ the set of $\mathcal{L}_{PA}$-formulae. We write $\forall t\varphi$ instead of $\forall x(ClTerm(x) \to \varphi)$ and $\forall v\varphi$ instead of $\forall x(Var(x) \to \varphi)$. Furthermore, we write e.g. $\forall tT\ulcorner\varphi(\underset{.}{t})\urcorner$ instead of $\forall x(ClTerm(x) \to T\underset{.}{s}(\ulcorner\varphi\urcorner,x))$. Again, this definition is extended to multi-variable cases in an obvious way. Then we can write a slightly stronger form of the uniform T-schema

$$\forall t_1 \ldots \forall t_n (T\ulcorner\varphi(\underset{.}{t_1},\ldots,\underset{.}{t_n})\urcorner \leftrightarrow \varphi(t_1^\circ,\ldots,t_n^\circ))$$

For more details on this notation, I refer the reader to Cantini [11] or Halbach [38].

We assume some standard coding for ordinals $< \Gamma_0$ and let $OT(x)$ represent the set (of codes) of ordinal terms. If α is an ordinal, we write $\overline{\alpha}$ for the numeral of its code. We write $\forall \alpha \varphi$ for $\forall x(OT(x) \to \varphi)$. We let $\prec$ represent the ordering of the ordinals in PA. PAT proves transfinite induction for every $\delta < \epsilon_0$, i.e. for all $\varphi \in \mathcal{L}_T$ and all $\delta < \epsilon_0$, PAT proves:

$$\forall \alpha(\forall \beta \prec \alpha \varphi(\beta) \to \varphi(\alpha)) \to \forall \zeta \prec \overline{\delta} \varphi(\zeta).$$

Unless otherwise specified, all axiomatic theories in this book are classical. Thus they are fully determined by specifying their non-logical axioms (and non-logical rules).

Standard models of $\mathcal{L}_T$ have the form $(\mathbb{N}, S)$, where $\mathbb{N}$ is the standard model of PA and $S \subseteq \omega$ interprets the truth predicate T. Let $Val_S(\varphi) = 1$ if and only if $(\mathbb{N}, S) \vDash \varphi$, where $\vDash$ is the classical satisfaction relation, and let $Val_S(\varphi) = 0$ otherwise. On occasion we also write φ^S for $Val_S(\varphi)$.

The most common way to compare axiomatic theories is by relative interpretations. Roughly, a theory $\mathcal{T}$ relatively interprets a theory $\mathcal{S}$ iff there is a translation from $\mathcal{L}_{\mathcal{S}}$ to $\mathcal{L}_{\mathcal{T}}$ that preserves logical structure of the formulae, possibly relativizing quantifiers, such that $\mathcal{T}$ proves the translations of all theorems of $\mathcal{S}$. The definition of a relative interpretation becomes more complicated if languages containing function symbols are considered; we omit an explicit definition and refer the reader to Halbach [38, ch. 6]. In this book, we will further demand that relative interpretations leave the arithmetical vocabulary untouched, with the possible exception of renaming of variables. This implies that if $\mathcal{T}$ relatively interprets $\mathcal{S}$, then all arithmetical theorems of the latter will also be provable in the former theory.

Fujimoto [29] has given a more fine-grained notion of interpretability in order to compare axiomatic theories *of truth*.

Assume that $\mathcal{S}$ and $\mathcal{T}$ are theories of truth (extending Peano arithmetic) formulated in the languages $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{T}}$ respectively. Assume that $\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{PA} \cup \{T_i | i \in I\}$, where $\{T_i | i \in I\}$ is the set of truth predicates of $\mathcal{L}_{\mathcal{S}}$ for some index set I. We say that $\mathcal{T}$ *defines the truth predicate(s) of* $\mathcal{S}$ iff for every $i \in I$ there is a formula $\varphi_i(x) \in \mathcal{L}_{\mathcal{T}}$ such that the result of uniformly substituting $\varphi_i(x)$ for T_i in a theorem of $\mathcal{S}$ is a theorem of $\mathcal{T}$.

More precisely, given a formula $\varphi_i(x)$ of $\mathcal{L}_{\mathcal{T}}$ for each $i \in I$, we define a function $h_{\vec{\varphi}}$ from $\mathcal{L}_{\mathcal{S}}$ to $\mathcal{L}_{\mathcal{T}}$ as follows:

$$h_{\vec{\varphi}}(\psi) = \begin{cases} \psi, & \text{if } \psi \text{ is an atomic formula of } \mathcal{L}_{PA} \\ \varphi_i(x), & \text{if } \psi = T_i(x) \\ \neg h_{\vec{\varphi}}(\chi), & \text{if } \psi = \neg \chi \\ h_{\vec{\varphi}}(\chi_1) \wedge h_{\vec{\varphi}}(\chi_2), & \text{if } \psi = \chi_1 \wedge \chi_2 \\ \forall x h_{\vec{\varphi}}(\chi), & \text{if } \psi = \forall x \chi \end{cases}$$

Then we say that $\mathcal{T}$ defines the truth predicate(s) of $\mathcal{S}$ iff there are formulae $\varphi_i(x)$ of $\mathcal{L}_T$ for each $i \in I$ such that $\mathcal{S} \vdash \psi$ implies $\mathcal{T} \vdash h_{\vec{\varphi}}(\psi)$ for all $\psi \in \mathcal{L}_{\mathcal{S}}$.

If $\mathcal{T}$ defines the truth predicate(s) of $\mathcal{S}$, then $\mathcal{T}$ relatively interprets $\mathcal{S}$. Since we assume that relative interpretations leave arithmetical vocabulary unchanged (except for renaming variables), this means $\mathcal{T}$ will prove all arithmetical theorems of $\mathcal{S}$.

2.2. Weak and strong diagonalization

All recursive functions are strongly represented in Peano arithmetic, but the language of Peano arithmetic (as we find it in most textbooks) does not contain function symbols for most of these functions. However, in investigating truth-theoretic axioms, one often works in a definitional expansion of Peano arithmetic. For example, in stating certain axioms it is often convenient to have in our language a symbol $\underset{\cdot}{\neg}$ for the function that sends the code of a sentence to the code of its negation. In the present section we will show that the choice of the base language really matters. More precisely, we will show that certain truth-theoretic axioms are inconsistent over PA when the language contains symbols for some primitive recursive functions, while they are consistent over PA when the language does not contain these function symbols amongst its vocabulary.

The usual way of achieving a self-referential sentence in the language of Peano arithmetic is by appeal to Gödel's diagonal lemma.

Proposition 2.2.1 (Diagonal lemma). *For every formula $\varphi(x)$ of $\mathcal{L}_{PA^-}$ with exactly x free, there exists a sentence ψ of $\mathcal{L}_{PA^-}$ such that $\mathsf{PA}^- \vdash \psi \leftrightarrow \varphi(\ulcorner\psi\urcorner)$.*

Here, PA^- is the theory of Peano arithmetic formulated in the language $\mathcal{L}_{PA^-}$ with signature $\{\overline{0}, S, +, \times\}$.

Proof. Let $f : \omega \to \omega$ by defined as follows. $f(n) = \#\varphi(\overline{n})$, if n is the code of $\varphi(x)$, and $f(n) = 0$ otherwise. Then f is recursive. Thus f is represented in PA^- by a formula $f^\circ(x, y)$. Now let some formula $\varphi(x)$ be given. Let θ be the formula $\exists x(f^\circ(y, x) \wedge \varphi(x))$, and let ψ be the formula $\theta(\ulcorner\theta\urcorner)$. Then $f(\#\theta) = \#\theta(\ulcorner\theta\urcorner) = \#\psi$. Hence PA^- proves $f^\circ(\ulcorner\theta\urcorner, \ulcorner\psi\urcorner)$. From this it follows that $\mathsf{PA}^- \vdash \psi \leftrightarrow \varphi(\ulcorner\psi\urcorner)$. □

Intuitively, it is the sentence on the left-hand side of the biconditional that is self-referential and not the one on the right-hand side. The right-hand side $\varphi(\ulcorner\psi\urcorner)$ refers only to ψ (i.e. mentions it), but not to $\varphi(\ulcorner\psi\urcorner)$. The sentence on the left-hand side refers to itself by way of definite description. For it is the only object that satisfies the formula $f^\circ(\ulcorner\theta\urcorner, x)$. One paragraph after introducing the sentence that now bears his name, Gödel writes: "We are therefore confronted with a proposition which asserts its own provability.” This remark is accompanied by the following footnote:

> In spite of appearances, there is nothing circular about such a proposition, since it begins by asserting the unprovability of a wholly determinate proposition [...], and only subsequently (and in some way by accident) does it emerge that this formula is precisley that by which the proposition was itself expressed. ([30, p. 41, fn 15])

Self-reference by way of definite description is therefore more similar to what people sometimes call contingent self-reference in natural language. (As an example, suppose that the only sentence written on the blackboard in room 223, Ludwigstr. 31, Munich, at 12 a.m. on August 31, 2014 is ‘The only sentence written on the blackboard in room 223, Ludwigstr. 31, Munich, at 12 a.m. on August 31, 2014 is false’.)

In his 2007 paper ‘Self-reference and the Language of Arithmetic’, Richard Heck [40] observes that there are some intuitively inconsistent principles of truth that are actually consistent in the standard language. He convincingly argues that true self-reference can only be achieved by expanding the standard language of arithmetic with function symbols for certain primitive recursive functions. Let us have a look at his example. Consider the following two truth-theoretic principles:

- $T\ulcorner\neg\varphi\urcorner \leftrightarrow \neg T\ulcorner\varphi\urcorner$ (Neg)
- $T\ulcorner Tt\urcorner \leftrightarrow Tt$ (T-Sym)

Heck provides a standard model for both principles (taken together), but argues that they should be inconsistent as follows. Suppose there were a term s such that $l = \ulcorner\neg Tl\urcorner$ is provable. This would be a formal representative of the ordinary Liar sentence

The Liar: The Liar is not true.

in our arithmetical language. Then we reach a contradiction as follows:

$$\begin{aligned} T\ulcorner Tl\urcorner &\leftrightarrow Tl, && \text{(T-Sym)} \\ &\leftrightarrow T\ulcorner\neg Tl\urcorner, && \text{substitution of identicals} \\ &\leftrightarrow \neg T\ulcorner Tl\urcorner, && \text{(Neg)} \end{aligned}$$

Terms like l become available once we enrich the language of arithmetic with functions symbols for certain primitive recursive functions (and appropriate axioms govering them). Let PA be the result of this expansion. Notice that PA conservatively extends PA^-. Then we get:

Proposition 2.2.2 (Strong diagonal lemma)**.** *For every formula $\varphi(x)$ of $\mathcal{L}_{PA}$ with exactly x free, there exists a term t of $\mathcal{L}_{PA}$ such that* $\mathsf{PA} \vdash t = \ulcorner\varphi(t)\urcorner$.

Proof. Given $\varphi(x)$, let $t := \underset{\cdot}{s}(\ulcorner\varphi(\underset{\cdot}{s}(x,x))\urcorner, \ulcorner\varphi(\underset{\cdot}{s}(x,x))\urcorner)$, where $\underset{\cdot}{s}$ is defined as in section 2.1. Now observe that

$$\underset{\cdot}{s}(\ulcorner\varphi(\underset{\cdot}{s}(x,x))\urcorner, \ulcorner\varphi(\underset{\cdot}{s}(x,x))\urcorner) = \ulcorner\varphi(\underset{\cdot}{s}(\ulcorner\varphi(\underset{\cdot}{s}(x,x))\urcorner, \ulcorner\varphi(\underset{\cdot}{s}(x,x))\urcorner)\urcorner = \ulcorner\varphi(t)\urcorner$$

□

Heck concludes "[t]rue self-reference is possible only if we expand the language to include function symbols for all primitive recursive functions. This language is therefore the natural setting for investigations of self-reference." ([40, p. 1])

The strong diagonal lemma seems to have made its first appearance in Jeroslow [48]. There he shows that one of Löb's derivability conditions can be dropped in the proof of Gödel's second incompleteness theorem once we work in the expanded language $\mathcal{L}_{PA}$.

The example presented by Heck is not an isolated case. Cain & Damnjanovic [9] have shown (fifteen years prior to the publication of Heck's paper) that the minimal Kripke fixed point under the Weak Kleene scheme (see section 3.2) is reached already after ω-many steps resp. only after ω_1^{CK}-many steps, depending on which Gödelcoding is chosen (and that therefore, the recursion-theoretic complexitiy of the fixed points depends essentially one the chosen Gödelcoding). Here is another example that I have found.

Proposition 2.2.3. *The scheme*

$$(\dagger) \quad T\ulcorner\neg Tt\urcorner \leftrightarrow \neg Tt$$

is inconsistent over PA.

Proof. By strong diagonalization, there is a term l such that

$$PA \vdash l = \ulcorner\neg Tl\urcorner.$$

Now we instantiate ($\dagger$) to l. Thus we get

$$T\ulcorner\neg Tl\urcorner \leftrightarrow \neg Tl.$$

By substitution of equals

$$T\ulcorner\neg Tl\urcorner \leftrightarrow \neg T\ulcorner\neg Tl\urcorner,$$

a contradiction. Hence ($\dagger$) is inconsistent over PA. □

Next we show the consistency of (†) over PA^-. Let $\mathcal{L}_T^-$ be the language of Peano arithmetic with signature $\{S, +, \cdot, \overline{0}\} \cup \{T\}$. The logical vocabulary comprises $\neg, \vee, \forall$ and $=$. All other connectives are defined in the usual way. Let $\langle n_1, \ldots, n_k \rangle := p_1^{n_1+1} \cdot \ldots p_k^{n_k+1}$, where p_i is the i-th prime number. Let $t^{\mathbb{N}}$ be the denotation of the term t in the standard model. Notice that the denotation function for terms of $\mathcal{L}_T^-$ is primitive recursive.

Let g be some p.r. Gödelcoding (for $\mathcal{L}_T^-$). Define the Gödelcoding g^+ by recursion as follows:

$$g^+(t) := \begin{cases} \langle t^{\mathbb{N}}, g(t) \rangle, \text{if } t \text{ is a closed term} \\ \langle 0, g(t) \rangle, \text{otherwise} \end{cases}$$
$$g^+(s = t) := \langle 0, g^+(s), g^+(t) \rangle$$
$$g^+(Tt) := \langle 1, g^+(t) \rangle$$
$$g^+(\neg\varphi) := \langle 2, g^+(\varphi) \rangle$$
$$g^+(\varphi \vee \psi) := \langle 3, g^+(\varphi), g^+(\psi) \rangle$$
$$g^+(\forall x \varphi) := \langle 4, g^+(x), g^+(\varphi) \rangle$$

Proposition 2.2.4. *1. $t^{\mathbb{N}} < g^+(t)$, where t is a closed term.*

2. $g^+(t) < g^+(Tt) < g^+(\neg Tt)$, where t is a closed term. Thus, it follows that there are no term fixed points under g^+.

Proposition 2.2.4 is immediate from the construction of g^+. Notice that g^+ is primitive recursive. Let e be an enumeration of $\{\neg Tt | t \text{ is a closed term}\}$ such that $i < j$ iff $g^+(e_i) < g^+(e_j)$. Let $\neg Tt_i := e_i$.

Proposition 2.2.5. *If $i \leqslant j$, then t_i cannot denote $\neg Tt_j$, i.e. $t_i^{\mathbb{N}} \neq g^+(\neg Tt_j)$.*

Proof. Case 1: $i = j$. By proposition 2.2.4 we have $t_i^{\mathbb{N}} < g^+(t_i) < g^+(\neg Tt_i)$.

Case 2: $i < j$. We have $g^+(\neg Tt_i) < g^+(\neg Tt_j)$ by definition of e. The claim follows because $t_i^{\mathbb{N}} < g^+(\neg Tt_i)$ by Proposition 2.2.4. □

Proposition 2.2.6. *The scheme (†) is consistent over* PA^-.

Proof. Let e and g^+ be as above.

Let $\Gamma_0 = \varnothing$.

Let $\Gamma_{i+1} = \begin{cases} \Gamma_i \cup \{g^+(\neg Tt_i)\}, \text{if } (\mathbb{N}, \Gamma_i) \vDash \neg Tt_i \\ \Gamma_i, \text{otherwise} \end{cases}$

Finally, let $\Gamma = \bigcup_{i \in \omega} \Gamma_i$

We show that $(\mathbb{N}, \Gamma) \vDash T\ulcorner \neg Tt_i \urcorner \leftrightarrow \neg Tt_i$ for all $i \in \omega$.

Assume (1) $(\mathbb{N}, \Gamma) \vDash \neg Tt_i$ in order to show (2) $(\mathbb{N}, \Gamma) \vDash T\ulcorner \neg Tt_i \urcorner$. Since the Γ_n are monotone, it follows that $(\mathbb{N}, \Gamma_n) \vDash \neg Tt_i$ for all n. In particular we have $(\mathbb{N}, \Gamma_i) \vDash \neg Tt_i$. Thus by definition $g^+(\neg Tt_i) \in \Gamma_{i+1} \subseteq \Gamma$. Thus (2) is proved.

Converse direction: Let (3) $(\mathbb{N},\Gamma) \vDash T\ulcorner\neg Tt_i\urcorner$ and assume for the sake of contradiction that (4) $(\mathbb{N},\Gamma) \vDash Tt_i$. From (3) and the definition of the Γ_j we conclude that (5) $(\mathbb{N},\Gamma_i) \vDash \neg Tt_i$. From (4) and the construction we conclude that $t_i^{\mathbb{N}} = g^+(\neg Tt_j)$ and $(\mathbb{N},\Gamma_j) \vDash \neg Tt_j$ for some $j < \omega$. Thus $g^+(\neg Tt_j) \in \Gamma_{j+1}$. But by Proposition 2.2.5 we have $j < i$, hence $j+1 \leqslant i$ and $\Gamma_{j+1} \subseteq \Gamma_i$ by construction. Thus $(\mathbb{N},\Gamma_i) \vDash T\overline{g^+(\neg Tt_j)}$. But this contradicts (5), since $t_i^{\mathbb{N}} = g^+(\neg Tt_j)$.

This completes the proof. □

3. Escaping the liar

Most philosophers view the T-schema as capturing something very important about the concept of truth. Disquotationalist make an even stronger claim: "the *basic* facts (i.e. the axioms of the theory that explains *every* other fact about truth) will all be instances of the above schema."[1] Let NT be the theory consisting of the axioms of PA plus all instances of the T-schema, $T\ulcorner\varphi\urcorner \leftrightarrow \varphi$, where φ is a sentence of $\mathcal{L}_T$. The acronym NT stands for 'naive truth'.

Proposition 3.0.7. *The theory* NT *is inconsistent.*

Proof. By the diagonal lemma, there is a sentence λ—a *liar* sentence—such that PA proves $\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$. By classical logic, the latter is equivalent to $\neg(T\ulcorner\lambda\urcorner \leftrightarrow \lambda)$. This contradicts the T-biconditional for λ, which is an axiom of NT. □

In order to block the derivation of the contradiction we are basically faced with two options. First, we can reject some inference rules of classical logic. It is known that weakening classical logic to intuitionistic logic is not enough (cf. Feferman [19])—so this option will have severe costs. Second, we can reject some instances of the T-schema. Both routes subdivide. An important subdivision of the second path is typing, which we discuss first.

3.1. Typing. Tarski's hierarchy

A very cautios way of restricting the T-schema, going back to Tarski, is to eschew all sentences that contain an occurrence of the truth predicate. This results in the theory TB (for 'Tarski-biconditionals', or sometimes DT for 'disquotational theory'). TB is the classical theory whose axioms are those of the base theory PAT plus all instances of the T-schema, $T\ulcorner\varphi\urcorner \leftrightarrow \varphi$, where φ is a sentence of the base language $\mathcal{L}_{PA}$ (i.e. a *T-free* sentence). It is not hard to prove that TB is consistent; in fact, it is conservative over PAT. TB is a *typed* theory of truth: it cannot prove the truth of a single sentence containing the truth predicate itself.[2]

[1] Horwich [47, p. 76]. Horwich actually does not talk about the T-schema but about the *equivalence scheme*—$\langle p\rangle$ is true iff p, where p ranges of propositions rather than sentences. This difference won't play a big role in our discussion.

[2] Cf. Halbach [38, chap. 10] for a short discussion of how to classify truth theories into typed and untyped (type-free) ones.

One of the main purposes of a truth theory is to facilitate the expression of generalizations. Does TB help us here? That question is taken up in Halbach [36] (see also Halbach [38, chap. 7]). His main result is the following:

Proposition 3.1.1 (Halbach). *Let $\varphi(x)$ be a T-free formula and let S be the set of all sentences of the form $\varphi(\ulcorner\psi\urcorner) \to \psi$, where ψ is a T-free sentence. Then the theories $S + \mathsf{PAT}$ and $\mathsf{TB} + \forall x(\varphi(x) \to Tx)$ have the same T-free consequences.*

Halbach remarks: "I take the result to be an exact formulation of the disquotationalist claim that infinite conjunctions can be expressed in a language containing a truth predicate which is characterized by the Tarskian equivalences. The infinite set $\varphi(\ulcorner\psi\urcorner) \to \psi$ of axioms replaces the infinite conjunction; therefore it can be avoided in order to introduce a formal system for a language comprising infinite conjunctions." ([36, p. 14]) Halbach's result shows that infinite conjunctions understood as sets of sentences of the form $\varphi(\ulcorner\psi\urcorner) \to \psi$ can be replaced by a single sentence of the form $\forall x(\varphi(x) \to Tx)$ in the presence of the restricted T-schema. Of course, TB only allows us to express infinite conjunctions of sentences that do *not* contain the truth predicate itself.

Proposition 3.1.1 can be strengthened. Let TO be the result of augmenting PAT with all instances of T-Out,

$$T\ulcorner\psi\urcorner \to \psi,$$

where ψ is again a sentence not containing the truth predicate. We will show that the above proposition still holds when the theory TB is replaced by the weaker theory TO. The observation is important in so far as it shows that, contrary to what most people might expect, the full T-biconditionals are not needed to express generalizations (at least so long as we understand 'express generalizations' in the way Halbach suggests): the left-to-right direction suffices.

Observation 3.1.2. *Let $\varphi(x)$ be a T-free formula and let S be the set of all sentences of the form $\varphi(\ulcorner\psi\urcorner) \to \psi$, where ψ is a T-free sentence. Then the theories $S + \mathsf{PAT}$ and $\mathsf{TO} + \forall x(\varphi(x) \to Tx)$ have the same T-free consequences.*

Proof. Clearly, if χ is a T-free consequence of $\mathsf{TO} + \forall x(\varphi(x) \to Tx)$, then χ is also a consequence of $\mathsf{TB} + \forall x(\varphi(x) \to Tx)$, because TO is a subtheory of TB. Thus, by proposition 3.1.1, χ is also a consequence of $S + \mathsf{PAT}$. Now let χ be a consequence of $S + \mathsf{PAT}$. Then only finitely many sentences in S have been used in the proof. Clearly, all of them follow from $\forall x(\varphi(x) \to Tx)$ plus the relevant instances of T-Out. □

A common complaint about TB, dating back to Tarski [89, p. 257], is that TB does not allow us to *prove* any non-trivial generalizations. For many philosophers, the deductive weakness has been a motivation to embrace a *compositional* theory truth such as CT. The theory CT is obtained by turning the inductive clauses of Tarski's truth definition into axioms.

Definition 3.1.3. The system CT is given by the axioms of PAT plus the following five axioms:

1. $\forall s \forall t(T(s \underset{\cdot}{=} t) \leftrightarrow s^\circ = t^\circ)$
2. $\forall x(Sent_{PA}(x) \rightarrow (T(\underset{\cdot}{\neg} x) \leftrightarrow \neg Tx))$
3. $\forall x \forall y(Sent_{PA}(x \underset{\cdot}{\wedge} y) \rightarrow (T(x \underset{\cdot}{\wedge} y) \leftrightarrow T(x) \wedge T(y)))$
4. $\forall x \forall y(Sent_{PA}(x \underset{\cdot}{\vee} y) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow T(x) \vee T(y)))$
5. $\forall x \forall v(Sent_{PA}(\underset{\cdot}{\forall} vx) \rightarrow (T(\underset{\cdot}{\forall} vx) \leftrightarrow \forall t T(x(t/v))))$

CT *does* prove certain generalizations; for example, it proves the global reflection principle for PA, i.e. the claim that all theorems of Peano arithmetic are true. The latter implies the consistency statement for PA. Therefore, by Gödel's second incompleteness theorem, CT is not conservative over PA. CT is actually much stronger than the consistency statement for PA: CT relatively interprets the second-order theory ACA.[3] The system ACA, in turn, is able to define the truth predicate of CT (cf. Takeuti [88]).

CT is still a typed theory of truth: the compositional axioms are restricted to sentences of the base language $\mathcal{L}_{PA}$. For example, CT proves that $1 = 1$ is true, but it does not prove that '$1 = 1$ is true' is true. At this point, we could either move to an untyped theory that allows us to apply the truth predicate to sentences containing the truth predicate. Or we stick to the Tarskian solution and introduce a *second* truth predicate, T_1, together with axioms that allow us to apply T_1 to sentences containing the original truth predicate T but that do not license the application of T_1 to sentences containing the new predicate. Such a theory would allow us e.g. to prove $T\ulcorner 1 = 1\urcorner$ and $T_1\ulcorner T\ulcorner 1 = 1\urcorner\urcorner$, but, again, not that the latter is true. In order to prove the latter, we might introduce a third truth predicate, T_2, with corresponding axioms that allow us to prove $T_2\ulcorner T_1\ulcorner T\ulcorner 1 = 1\urcorner\urcorner\urcorner$. There is, of course, no need to stop here. We can introduce a predicate T_n for every natural number n; and once we have done this, we can iterate this construction into the transfinite. Of course, since we want to be able to reason about the syntax of this theory within PA, we should only expand this hierarchy along some computable well-ordering and probably only along well-orderings whose well-foundedness can be proved within the base theory (or within any of the theories that we have already added on top of the base theory).

[3] A definition of ACA can be found in the Appendix. For a proof of the claim we refer the reader to Halbach [38].

In chapter 8 we will consider Tarski hierarchies of height ω_1^{CK} (=the least nonrecursive ordinal, called the 'Church-Kleene ordinal'), but in our proof-theoretic investigations we will only consider hierarchies of height (at most) Γ_0 (=the least strongly critical ordinal, called the 'Feferman-Schütte ordinal'). More details on the coding of ordinals can be found in the appendix. The formalization of Tarski's hierarchy presented below is, to my knowledge, due to Halbach [34]. The acronym RT stands for 'ramified truth'. Let $\mathcal{L}_T^\gamma$ be the language of Peano arithmetic augmented with truth predicates T_α for every $\alpha \prec \gamma$, and let $Sent_\gamma(x)$ be a formula that naturally represents the sentences of $\mathcal{L}_T^\gamma$.

Definition 3.1.4. The system RT_γ is given by the axioms of PA with full induction in the language $\mathcal{L}_T^\gamma$ plus the following axioms, for all $\alpha \prec \gamma$:

1. $\forall s \forall t (T_\alpha(s \underset{.}{=} t) \leftrightarrow s^\circ = t^\circ)$
2. $\forall x (Sent_\alpha(x) \to (T_\alpha(\underset{.}{\neg} x) \leftrightarrow \neg T_\alpha x))$
3. $\forall x \forall y (Sent_\alpha(x \underset{.}{\land} y) \to (T_\alpha(x \underset{.}{\land} y) \leftrightarrow T_\alpha(x) \land T_\alpha(y)))$
4. $\forall x \forall y (Sent_\alpha(x \underset{.}{\lor} y) \to (T_\alpha(x \underset{.}{\lor} y) \leftrightarrow T_\alpha(x) \lor T_\alpha(y)))$
5. $\forall x \forall v (Sent_\alpha(\underset{.}{\forall} v x) \to (T_\alpha(\underset{.}{\forall} v x) \leftrightarrow \forall t T_\alpha(x(t/v))))$
6. $\forall t (Sent_\beta(t^\circ) \to (T_\alpha(\underset{.}{T}_\beta(t)) \leftrightarrow T_\beta(t^\circ)))$ for $\beta \prec \alpha$
7. $\forall t \forall \beta \prec \overline{\alpha} (Sent_\beta(t^\circ) \to (T_\alpha(\underset{.}{T}_\beta(t)) \leftrightarrow T_\alpha(t^\circ)))$

The theory CT is just RT_1. RT has some proof-theoretic power. Let RA_α be the system of ramified analysis up to level α (cf. Feferman [18]).

Theorem 3.1.5 (Feferman [20]). RT_α *and* RA_α *are proof-theoretically equivalent.*

The systems RA_α are 'semi-formal' systems, containing infinitary limit generalization rules, and are therefore not very attractive. Halbach [34] has given a match up between the systems RA_α and the systems $(\Pi_1^0 - \mathsf{CA})_{\omega \cdot \alpha}$.

The Tarskian solution of the paradoxes has been critizised on several grounds. One complaint is that typing is *ad hoc* and only motivated by the desire to evade the paradoxes. A second objection is that truth is a univocal concept that is fragmented in the Tarskian approach, or that typing (indexing) is not found in natural language. I do not find these arguments very compelling. As Carnap has taught us, "The explicatum is to be *similar to the explicandum* in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted" ([12, p. 7]), namely when the divergence is justified by the fruitfulness and simplicity of the explicatum. An explication or regimentation does not have to respect all the features that an idiom exhibits in natural language. Naturalness is something that we cherish, of course, but as long the solution is fruitful and satisfies our needs, we should be content with it. (Moreover, intuitions about the paradoxical statements differ widely.)

The problem with the Tarskian solution, then, is not that it is artifical or *ad hoc*; the problem is that typed truth predicates simply do *not* satisfy our needs. RT_{Γ_0} does not allow us to express certain generalizations that we would like to express. Let $\Phi \subseteq \mathcal{L}_T^{\Gamma_0}$ be some set of theorems of RT_{Γ_0}, represented in PA by the formula $\varphi(x)$; and assume that the truth predicates (i.e. their indices) occurring in Φ are unbound in Γ_0. (For example, let Φ be the set of all sentences of the form $\psi \to \psi$.) Then no matter which index α we choose, the sentence $\forall x(\varphi(x) \to T_\alpha(x))$ will fail to express the infinite conjunction of the members of Φ. But one of the main reasons why we want to have a truth predicate in the first place is our desire to capture infinite sets of sentences by a single sentence. In particular, we would like to express agreement with our own theory. In order to formulate a global reflection principle for the Tarskian hierarchy of, say, level α, we have to move one level up in the hierarchy. Of course, we can go still a bit beyond Γ_0, but there will be a point at which the hierarchy becomes unmanagable (if we want a *boundedly recursive* hierarchy (cf. Appendix), i.e. a hierarchy along a well-ordering such that all ots initial segments are recursive, then ω_1^{CK} is the halting point).

3.2. Non-classical solutions. Kripke fixed points

The most influential approach to break the binds of the Tarskian hierarchy is Kripke's 'Outline of a theory of truth' [53], to which we now turn. Although we will argue that non-classical solutions are ultimately unsatisfactory, we present Kripke's approach in some detail because his models will be very helpful in providing an analysis of the paradoxes, as a guidance in devising axiomatic truth theories, and in giving consistency proofs for them.

Kripke follows the widely shared view that the liar sentence does not succeed in expressing a proposition, or lacks a definite truth value. If we do not ban such sentences from our language, we need a framework for reasoning with them. Partial models provide such a framework.

Definition 3.2.1. A *partial model* for $\mathcal{L}_T$ is a pair $S = (S^+, S^-)$, where $S^+, S^- \subseteq Sent_T$. S^+ is called the *extension* and S^- is called the *anti-extension* of the truth predicate under S. We write (slightly abusing notation) $S_1 \subseteq S_2$ iff $S_1^+ \subseteq S_2^+$ and $S_1^- \subseteq S_2^-$. A partial model is *consistent* iff $S^+ \cap S^- = \varnothing$. A *partial valuation* is a function from $Sent_T \to \{0, 1, \frac{1}{2}\}$. A *valuation scheme* V is a function from partial models to partial valuations.

Let us introduce some of the valuation schemes that will be important later on. The *Strong Kleene* valuation scheme V_{SK} is defined by induction as follows.

1. $V_{SK}(S)(s=t) = \begin{cases} 1, & \text{if } s^{\mathbb{N}} = t^{\mathbb{N}} \\ 0, & \text{if } s^{\mathbb{N}} \neq t^{\mathbb{N}} \end{cases}$

2. $V_{SK}(S)(Tt) = \begin{cases} 1, & \text{if } t^{\mathbb{N}} \in S^+ \\ 0, & \text{if } t^{\mathbb{N}} \in S^-\text{or } t^{\mathbb{N}} \notin Sent_T \\ \frac{1}{2}, & \text{otherwise} \end{cases}$

3. $V_{SK}(S)(\neg\varphi) = 1 - V_{SK}(S)(\varphi)$

4. $V_{SK}(S)(\varphi \wedge \psi) = min\{V_{SK}(S)(\varphi), V_{SK}(S)(\psi)\}$

5. $V_{SK}(S)(\forall x\varphi) = min\{V_{SK}(S)(\varphi(t/x)) | t \text{ is a closed term}\}$

Under the Strong Kleene scheme, the value $\frac{1}{2}$ can be interpreted as 'indeterminate' or 'unknown'.

The *Weak Kleene* valuation scheme V_{WK} is defined exactly as the Strong Kleene scheme, except that one uses the non-standard order $(\frac{1}{2}, 0, 1)$ for the computation of the minimum in the clauses for the conjunction and the quantifier. Under the Weak Kleene scheme, the value $\frac{1}{2}$ is best understood as 'meaningless' or 'non-sense'.

Call a set $P \subseteq \omega$ *consistent* iff $\#\varphi \in P$ implies $\#\neg\varphi \notin P$. The *supervaluational* valuation scheme V_{FV} is defined as follows.

$$V_{FV}(S)(\varphi) = \begin{cases} 1, & \text{if for all consistent } P \supseteq S^+ : (\mathbb{N}, P) \vDash \varphi \\ 0, & \text{if for all consistent } P \supseteq S^+ : (\mathbb{N}, P) \vDash \neg\varphi. \\ \frac{1}{2}, & \text{otherwise} \end{cases}$$

A valuation scheme V is *monotonic* iff for all partial models S_1, S_2 with $S_1 \subseteq S_2$ we have: if $V(S_1)(\varphi) = 1$, then $V(S_2)(\varphi) = 1$ and if $V(S_1)(\varphi) = 0$, then $V(S_2)(\varphi) = 0$. All of the valuation schemes introduced above are monotonic. The classical valuation *Val*, on the other hand, is not.

Definition 3.2.2. Let S be a partial model and V a valuation scheme. The *Kripke-jump of* S (relative to V) is defined as follows: $\mathcal{J}_V(S) = (\mathcal{J}_V(S)^+, \mathcal{J}_V(S)^-)$, where

$$\mathcal{J}_V(S)^+ = \{\#\varphi | V(S)(\varphi) = 1\}$$
$$\mathcal{J}_V(S)^- = \{\#\varphi | V(S)(\varphi) = 0\} \cup \{n | n \notin Sent_T\}$$

In the Kripke-jump of S, every sentences that receicves value 1 in S is declared true, and every sentences that receives value 0 in S is declared untrue.

Theorem 3.2.3 (Kripke). *If V is a monotonic valuation scheme then the operator $\mathcal{J}_V$ is monotone, i.e. $S_1 \subseteq S_2$ implies $\mathcal{J}_V(S_1) \subseteq \mathcal{J}_V(S_2)$.*

This is proved by induction on the complexity of formulae. The monotonicity of $\mathcal{J}_V$ implies that there are fixed points, i.e. partial models S with $S = \mathcal{J}_V(S)$. This follows from a simple cardinality argument. These fixed points can be 'reached from below' as follows:

Definition 3.2.4. For each ordinal α and partial model $S = (S^+, S^-)$ we inductively define the partial model $\mathcal{J}_V^\alpha(S)$ by transfinite recursion as follows.

1. $\mathcal{J}_V^0(S) = S$
2. $\mathcal{J}_V^{\alpha+1}(S) = \mathcal{J}_V(\mathcal{J}_V^\alpha(S))$
3. $\mathcal{J}_V^\gamma(S) = (\bigcup_{\alpha<\gamma} \mathcal{J}_V^\alpha(S)^+, \bigcup_{\alpha<\gamma} \mathcal{J}_V^\alpha(S)^-)$, if γ is a limit ordinal

Call a partial model S *sound* iff $S \subseteq \mathcal{J}_V(S)$. The partial model $(\varnothing, \varnothing)$ is trivially sound.

Theorem 3.2.5 (Kripke)**.** *If V is a monotonic valuation scheme and S is sound then there is an α such that $\mathcal{J}_V^\alpha(S) = \mathcal{J}_V^{\alpha+1}(S)$. We denote this fixed point by $\mathcal{J}_V^\infty(S)$.*

The *minimal* Kripke fixed point of $\mathcal{J}_V$ is the pair $\mathcal{J}_V^\infty((\varnothing, \varnothing))$, which (by slight abuse of notation) we simply denote by $\mathcal{J}_V^\infty(\varnothing)$. Accordingly, the extension (anti-extension) of the truth predicate in the minimal fixed point is denoted by $\mathcal{J}_V^\infty(\varnothing)^+$ $(\mathcal{J}_V^\infty(\varnothing)^-)$.
Much of the interest that we have in fixed-point models derives from the following property:

Theorem 3.2.6 (Kripke)**.** *If $S = \mathcal{J}_V(S)$, then*

$$V(S)(T\ulcorner\varphi\urcorner) = V(S)(\varphi)$$

for all $\varphi \in \mathcal{L}_T$.

In a fixed point, the sentences with value 1 are exactly those sentences that are in the extension of the truth predicate and the sentences with value 0 are exactly those sentences that are in the anti-extension of the truth predicate. In a certain sense, then, Kripke provides us with a model theory for languages, based on a non-classical logic, that can represent their own truth predicate.

Besides that, Kripke also gave very useful definitions of groundedness and paradoxicality. We will study these concepts extensively in Part II of this book.

Definition 3.2.7. A sentence φ is *grounded* (relative to V) iff φ has a definite truth value in the minimal fixed point of $\mathcal{J}_V$, i.e. iff $V(\mathcal{J}_V^\infty(\varnothing))(\varphi) \in \{0, 1\}$. A sentence φ is *Kripke-paradoxical* (relative to V) iff there is no fixed point S of $\mathcal{J}_V$ such that φ has a definite truth value in S.

Definition 3.2.8. A valuation scheme V is *classically sound* iff for all consistent partial models $S = (S^+, S^-)$ and all sentences φ the following holds: if $V(S)(\varphi) \in \{0, 1\}$, then $V(S)(\varphi) = Val_{S^+}(\varphi)$.

The classical model $(\mathbb{N}, \mathcal{J}_V^\infty(S)^+)$ is called the 'close off' of the partial model $\mathcal{J}_V^\infty(S)$.

Proposition 3.2.9. *If V is monotone and classically sound and S is a sound partial model, then*

$$(\mathbb{N}, \mathcal{J}_V^\infty(S)^+) \vDash T\ulcorner\varphi\urcorner \leftrightarrow \varphi$$

for all $\varphi \in S^+ \cup S^-$.

Proof. This follows from Theorem 3.2.6 and the definition of a classically sound evaluation scheme. □

In particular, Proposition 3.2.9 implies that the T-schema for all grounded sentences is ω-consistent in classical logic. We will return to that later.

Philosophers have paid much attention to the 'internal' theory of the minimal Strong Kleene fixed point, KFS ([26], [86]). KFS is the theory consisting of all the sentences that have value 1 in the minimal Strong Kleene fixed point, i.e. KFS $= \mathcal{J}_{SK}(\varnothing)^+$. It is a *paracomplete* theory; it is based on the logic K_3 in which the law of exluded middle does not hold. By the compositionality of the Strong Kleene scheme and Theorem 3.2.6, KFS satisfies the Intersubstitutivity Principle.

KFS has been critizised for several reasons. First, although it satisfies the Intersubsitutivity Principle, it does not satisfy the unrestricted T-schema. In fact, it does not satisfy any of its two directions: From $T\ulcorner\varphi\urcorner \rightarrow \varphi$ we would get, by Intersubstitutivity, $\varphi \rightarrow \varphi$. But this is equivalent to $\neg\varphi \vee \varphi$, which does not hold in K_3.[4] For the same reason, $\varphi \rightarrow T\ulcorner\varphi\urcorner$ does not hold. But keeping the unrestricted T-schema seems to be the *raison d'être* for abandoning classical logic. A second weakness of KFS is that it lacks a decent conditional. As already remarked, we don't have $\varphi \rightarrow \varphi$ in K_3. In Feferman's [19] words, "nothing like sustained ordinary reasoning can be carried on in [this] logic." Third, KFS is unable to express that the liar sentence is 'gappy' (neither true nor false) or otherwise defective. Fourth, KFS is not a theory in the true sense of the word. From a recursion-theoretic point of view, the theory KFS is very complex:[5]

Theorem 3.2.10 (Kripke, Burgess [8])**.** *The set $\mathcal{J}_{SK}^\infty(\varnothing)^+$ is Π_1^1-complete.*

[4] Notice that supervaluational fixed points do *not* satisfy Intersubstitutivity, albeit they satisfy Theorem 3.2.6! This is because all classical tautologies hold under the supervaluational scheme.

[5] See the appendix for some background on recursion theory. A proof of the result is given in chapter 7.

The theory KFS is therefore not recursively axiomatizable. There is now a lively research activity, initiated by Hartry Field [26], that tries to improve KFS by adding a conditional that is not defined in terms of negation and disjunction (these approaches are even more complex than Kripke's fixed points). I won't go into the details of these approaches, but sketch three reasons why I think that non-classical truth theories are ultimately unsatisfactory.

One difference between classical and non-classical truth theorists is with respect to the acceptance of the T-schema. Classical truth theorists have to reject—in fact, accept the negation of—certain instances of the T-schema: if λ is a liar sentence, then $\neg(T\ulcorner\lambda\urcorner \leftrightarrow \lambda)$ will already be provable in PA. Non-classical truth theorists, on the other hand, will accept the unrestricted T-schema—keeping the unrestricted T-schema (or the general equivalence between a sentence and its truth predication) is the *raison d'être* for abandoning classical logic. One reason against adopting all instances of the T-schema is the so-called *revenge* phenomenon.[6]

Consider, again, the liar paradox, i.e. a sentence that says of itself that it is not true. A common response to the antinomy is to declare that the liar sentence is neither true nor false. But then there is an obvious problem: since the liar is neither true nor false, it is in particular not true. But this is just what the liar sentence says. Thus, if we accept the equivalence of any sentence and its truth predication, then the assumption that the liar sentence is neither true nor false leads us to the conclusion that the liar sentence is true after all. A truth theory based on rejection of bivalence is therefore either bound to be inconsistent or it won't be able to express the defectiveness of the liar within the object language. Now, a common reaction is to introduce a new predicate—say, 'defective'—and declare that the liar falls under that predicate. But then a new paradox will emerge: consider a sentence that says of itself that it is either untrue or defective. The latter is a so-called revenge-liar. The existence of revenge liars is usually taken to show that the original liar paradox has not been solved properly. As far as I can see, the revenge phenomenon is a problem that pertains *exclusively* to non-classical solutions: it is generated by the adherence to the general equivalence of a sentence and its truth predication. There is no problem with saying 'The liar is not true', unless this forces me to say "The liar is not true' is true'. And I am only forced to say the latter if I adhere to the unrestricted equivalence of a sentence and its truth predication. Classical truth theorists, on the other hand, do not face this problem. They can consistently declare the liar to be neither true nor false, using the very same vocabulary which was used to formulate the liar sentence. And the reason why the assumption that the liar sentence is neither true nor false won't lead to a contradiction lies simply in the fact that by rejecting the T-biconditional for the liar, the classical truth theorist is not forced to conclude that the liar is thereby true after all.

[6] For more on the revenge phenomenon, see the essay collection Beall [5].

Here is another argument against non-classical truth theories. The main reason for having a truth predicate in the first place is that we want to increase our expressive power—and weakening classical logic seems to have quite the opposite effect. For example, all theories in which the unrestricted T-schema holds are able to derive all comprehension axioms of second-order arithmetic—in fact, all axioms of n-th order arithmetic, for arbitrarily large n (by Russell's trick, cf. section 1.2 and Part III of this thesis). However, though derivable, these comprehension axioms cannot be used to establish (many) facts about arithmetic. For example, the model-theoretic constructions underlying most non-classical approaches can already be carried out in proper subsystems of second-order arithmetic. (For example, the minimal Strong Kleene fixed point can be defined using only Π^1_1-comprehension.) But if these theories are consistent, they cannot prove their own consistency, by Gödel's second incompleteness theorem. Thus, although these theories can prove quite a few comprehension axioms, their underlying logic does not allow them to reason properly with them. We will see later that there are classical truth theories, based on some subset of the T-schema, that do prove the consistency of these non-classical theories. Although such theories contain fewer instances of the T-schema than the non-classical ones, and thereby fewer instances of comprehension, they can prove more arithmetical facts because they can reason classically with these comprehension axioms.

Of course, one might object that the expressive power gained by a truth predicate should not be measured (solely) on the basis of their proof-theoretic strength. For example, the Tarskian hierarchy RT_{Γ_0} is remarkable in its deductive power (cf. Theorem 3.1.5), but it does not allow us to express all the generalizations that we want: if Φ is a set of theorems the truth indices of which are unbound in Γ_0, then we cannot express (within RT_{Γ_0}) that all members of Φ are true. So the proponents of non-classical solution will claim that they fare better at expressing generalizations.

This brings us to the most important point. The truth predicate is used in many areas of philosophy. If truth obeys non-classical laws, non-classicality will spread over to other concepts that are defined in terms of truth, e.g. knowledge.[7] Although Quine [70] has taught us that no sentence is immune to revision, it is clear that we should not revise the basis of our web of believe unless there is no other way out. The main reason for having a truth predicate at all is that it enables us to express generalizations that we could not express otherwise, namely, generalizations that serve as a proxy for infinite conjunctions. Quite a few authors have argued that the unrestricted T-schema is a necessary condition in order to fulfill this function: this seems to be the main argument against classical truth theories. But if the generalizing function could be fulfilled in classical logic, there would be no reason at all why we should move to a non-classical logic. We have already seen (Observation

[7] This example has in particular been stressed by Volker Halbach.

3.1.2), in the case of the typed theory TB, that only one direction of the T-schema is needed in order to express infinite conjunctions. As we will show in the next chapter, a similar point can be made when one considers untyped theories of truth. The argument that the T-schema is necessary for the generalizing function of truth is unsound, and thus the main argument against classical truth theories breaks down. There is no reason for weakening classical logic.

4. Classical untyped truth

A classical truth predicate cannot validate the Intersubstitutivity Principle nor the unrestricted T-schema, on pain of contradiction. This, however, is on itself hardly a convincing argument against classical truth theories. What we demand from the logician is a regimentation (in the Quinean sense) of a certain fragment of our language. Such a regimentation does not have to respect all the features that an idiom exhibits in ordinary language. On that view, whether the concept of truth is "at bottom disquotational" or whether the T-biconditionals "fix its meaning" is only of secondary importance—in particular for deflationists that do not grow tired of claiming that "the truth predicate exists solely for a certain logical need" (Horwich [46, p. 2]).

The job of the logician is to provide us with a device that serves our needs—a device that allows us to express agreement or disagreement with an infinite bunch of sentences, that enables us to express our commitment to a theory like PA, etc. A convincing argument against classical truth theories would have to show that this is *not* possible in classical logic.

Hartry Field ([24], [25], [26]) has given arguments that are supposed to establish that classical truth predicates cannot fulfill the functions that we want it to fulfill. In this chapter we will try to refute these arguments.

4.1. Expressing infinite conjunctions

It is common (cf. Armour-Garb [1]) to distinguish between the *expressibility* and the *provability* of generalizations (viewed as infinite conjunctions). Any decent theory of truth should at least allow for the expression of generalizations; whether such generalizations ought to be provable too is a more difficult question. Clearly, there are generalizations that we do not expect a theory of truth to prove—e.g. 'Everything that Einstein said about relativity is true'. Moreover, there are generalizations that we certainly do not want our theory to prove—e.g. 'All sentences of the form '$A \to \bot$' are true'. But even if we reject such generalizations, we might nevertheless want to be able to assert them hypothetically in the course of some argument. Therefore, we will focus on the task of expressing generalizations.

4.1.1. What does it take to express an infinite conjunction?

The generally accepted picture of how the notion of truth allows us to express generalizations or 'infinite conjunctions' is nicely embodied in the following passage of Horwich ([46, p. 3]):

> Consider, for example,
>
> (1) What Oscar said is true.
>
> Here we have something of the form
>
> (2) x is F,
>
> whose meaning is such that, given further information about the identity of x—given further premises of the form
>
> (3) x =the proposition that p
>
> —we are entitled to infer
>
> (4) p.
>
> And it is from precisely this inferential property that propositions involving truth derive their utility. For it makes them, in certain circumstances, the only appropriate object of our beliefs, suppositions, desires, etc. Suppose, for example, I have great confidence in Oscar's judgment about food; he has just asserted that eels are good but I didn't quite catch the remark. Which belief might I reasonably acquire? Well, obviously not that eels are good. Rather what is needed is a proposition from which that one would follow, given identification of what Oscar said—a proposition equivalent to
>
> If what Oscar said is *that eels are good* then eels are good, and if he said *that milk is white* then milk is white, ... and so on;
>
> and the *raison d'etre* of the concept of truth is that it supplies us with such a proposition: namely (1).

Thus, according to Horwich and many other deflationists, a generalization such as 'What Oscar said is true' is, by virtue of the disquotational nature of truth, somehow *equivalent* to the infinite conjunction

$$\bigwedge_{\psi\in\mathcal{L}} \text{If Oscar said } \ulcorner\psi\urcorner \text{ then } \psi,$$

and therefore *implies*—given an identification of what Oscar said—the proposition asserted by Oscar, namely, that eels are good. And it is this *inferential property* that makes generalizations involving the truth predicate *useful*. Of course, this purported explanation raises a couple of important questions that we should clearly separate from each other:

1. In what sense can a generalization

   $$\forall x(\varphi(x) \to Tx) \tag{4.1}$$

   and an infinite conjunction

   $$\bigwedge_{\psi \in \mathcal{L}} (\varphi(\ulcorner \psi \urcorner) \to \psi) \tag{4.2}$$

   be said to be equivalent (relative to some theory of truth)? Or: In what sense can a generalization express an infinite conjunction?

2. What conditions does a theory of truth have to satisfy in order to yield this equivalence? In particular: Does the equivalence presuppose the T-schema or the Intersubstitutivity Principle?

3. Does the equivalence between generalizations and infinite conjunctions yield the inferential properties from which generalizations derive their utility? Do the inferential properties from which generalizations derive their utility presuppose the equivalence?

In [36], Halbach attempts to answer the first of these questions as follows. First, the infinite conjunction (4) is replaced by the infinitely many sentences

$$\varphi(\ulcorner \psi \urcorner) \to \psi \tag{4.3}$$

(Here, in order to avoid problems with liar-like sentences, Halbach considers only sentences ψ and predicates $\varphi(x)$ that do *not* contain the truth predicate.) Now, given all instances of the (typed) Tarski-biconditionals, all instances of (5) follow from (3). Of course, there are many other sentences which imply all instances of (5); every contradiction does the job. However, as Halbach observes, (3) only implies sentences in the truth-free language that are implied by the instances of (5). That is, given the typed Tarski-biconditionals, (3) has *exactly* the same truth-free consequences as the infinitely many premises (5). (Cf. Proposition 3.1.1.) Generalising on Halbach's observation, we start our discussion with the following proposal.

Definition 4.1.1 (temporarily)**.** Let Γ be some (typed) theory of truth extending PA. We say that Γ *enables us to express infinite conjunctions* over truth-free sentences if, and only if, for every T-free predicate $\varphi(x)$, the theory

$$\mathsf{PA} + \{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi | \psi \text{ is T-free}\}$$

has exactly the same T-free consequences as the theory

$$\Gamma + \forall x(\varphi(x) \rightarrow Tx)$$

In the above definition, Γ might be any (typed) theory of truth, disquotational or otherwise. Γ does not have to be an axiomatic theory of truth but might be a semantic theory of truth, i.e. a set of sentences true in some model, and therefore does not have to be recursively enumberable. Moreover, the consequence relation might be non-classical or given by some restricted class of models and therefore not effective (as is the case with the revision theory of truth or Kripke's family of fixed-point models).

Given this definition, we can answer our three questions (partly) as follows. First, generalizations and infinite conjunctions are equivalent with respect to their truth-free consequences. Second, as is easily seen, the typed Tarski-biconditionals are sufficient for ensuring this equivalence (at least in classical logic). (We will point out a necessary condition below.) Whether all uses of generalizations can be explained on the basis of their equivalence with infinite conjunctions can only be answered if we are given a complete list of all their uses, but we can see that at least some of the most important uses are explained. If Γ is recursively enumerable and the consequence relation is effective, then generalizations allow us to *finitely axiomatize* the infinitely many premises with a single sentence (relative to the truth theory Γ, anyways).

Unfortunately, the above definition has some counterintuitive consequences. Observe that, according to the above definition, any truth theory that enables us to express infinite conjunctions must be *conservative* over its base theory. That is, if φ is a T-free sentence that follows from the truth theory Γ then φ must already be a consequence of the base theory PA. Although some authors (e.g. Horsten [44], Shapiro [84], Ketland [49]) have argued that a deflationary truth theory ought to be conservative over its base theory, this is controversial. But even if a deflationary theory ought to be conservative over its base theory, it should not be implied by a criterion on the expressibility of infinite conjunctions.

First of all, a good criterion should also work for non-deflationary or substantial theories of truth. It is not clear why such theories ought to satisfy some conservativeness requirement. Substantial truth theorists do not necessarily deny that truth has expressive functions; they argue that truth is not *merely* an expressive device.

And a good definition should also make us understand under which conditions a substantial theory of truth allows us to express infinite conjunctions.

Secondly, and more importantly, if Γ is a truth theory that enables us to express infinite conjunctions and Γ^* is a truth theory that extends Γ, then intuitively Γ^* should *also* enable us to express infinite conjunctions. However, the conservativity restriction precludes this. For example, consider the (typed) compositional theory of truth, CT. This theory contains TB as a subtheory. According to Proposition 3.1.1, the theory TB enables us to express infinite conjunctions. But then we should expect that CT enables us to do the same, because it contains TB. However, the theory CT is not conservative over its base theory. Thus, according to definition 4.1.1, the compositional theory of truth does not enable us to express infinite conjunctions. We may try to remedy this defect by replacing the occurrence of 'PA' in the definition by 'Γ'. Thus:

Definition 4.1.2 (temporarily)**.** Let Γ be some (typed) theory of truth. We say that Γ *enables us to express infinite conjunctions* over truth-free sentences if, and only if, for every T-free predicate $\varphi(x)$, the theory

$$\Gamma + \{\varphi(\ulcorner\psi\urcorner) \to \psi | \psi \text{ is T-free}\}$$

has exactly the same T-free consequences as the theory

$$\Gamma + \forall x(\varphi(x) \to Tx)$$

However, although this definition does not imply that the truth theory is conservative over its base theory, it is still possible to find pairs of theories $\Gamma \subseteq \Sigma$ such that Γ satisfies the criterion but Σ doesn't. Here is an example. Let Γ be TO and let Σ be $\mathsf{TO} + \neg T\ulcorner\chi\urcorner$, where χ is some arithmetical statement such that neither χ nor $\neg Prov(\ulcorner\chi\urcorner)$ are derivable from $\mathsf{TO} + \{Prov(\ulcorner\psi\urcorner) \to \psi | \psi \in \mathcal{L}\}$. It follows from proposition 3.1.2 that TO satisfies definition 4.1.2. However, Σ doesn't. Observe that $\mathsf{TO} + \neg T\ulcorner\chi\urcorner + \forall x(Prov(x) \to Tx)$ proves $\neg Prov(\ulcorner\chi\urcorner)$ while $\mathsf{TO} + \neg T\ulcorner\chi\urcorner + \{Prov(\ulcorner\psi\urcorner) \to \psi | \psi \text{ is T-free}\}$ doesn't.

Of course, $\mathsf{TO}+\{Prov(\ulcorner\psi\urcorner) \to \psi | \psi \text{ is T-free}\}$ proves all instances of $\forall x(Prov(x) \to Tx)$, but a generalization usually has *more* force than its instances, so we cannot expect them to have the same (T-free) consequences. We may attempt to improve the last definition by proposing that the two theories should have the same consequences *given some infinitary rule* such as the ω-rule. Some semantic theories of truth are actually closed under the ω-rule. For example, it is not hard to see that over Kripke's theory with the Strong Kleene valuation scheme, generalizations and infinite conjunctions (understood as an infinite set of premises) have the same consequences (have the same truth conditions). Moreover, it is easily seen that axiomatic theories like TB or CT satisfy the equivalence between generalizations and infinite conjunctions if the equivalence is interpreted as equivalence with respect to their consequences given the ω-rule. But while the last proposal seems to provide a plausible answer to our first question—in what sense can we say that an infinite conjunction and a generalization are equivalent?—it is by no means clear why a theory satisfying the definition would be useful in *actual* reasoning. For finite reasoners like us, the fact that generalizations are equivalent to infinite conjunctions with respect to their consequences in some infinitary logic (or over the standard model of arithmetic) seems rather futile. We can neither employ the ω-rule in actual reasoning nor do we have a full grasp of the standard model of arithmetic.

There are two possible responses. The first one is to reject the proposed explication of the equivalence between generalizations and infinite conjunctions. The second one is to conclude that the inferential properties that make a generalization useful in actual reasoning do not presuppose the *full* equivalence between generalizations and infinite conjunctions. In order to clarify this, we would like to propose another explication of the alleged equivalence between generalizations and infinite conjunctions. The idea, this time, is that they should be equivalent *with respect to their inferential behavior.*

Deflationists claim that the truth predicate is a quasi-logical device, comparable to a connective. The use of a connective is commonly characterized by introduction and elimination rules. In the case of infinite conjunctions, the following two rules are the most natural candidates: to infer A_i from $\bigwedge_{j\in I} A_j$ (elimination) and to infer $\bigwedge_{j\in I} A_j$ from $\{A_i | i \in I\}$ (introduction). We will discuss both rules in turn, starting with the elimination rule. Given the elimination rule for infinite conjunctions, we can infer $\varphi(\ulcorner\psi\urcorner) \to \psi$ from $\bigwedge_{\psi\in\mathcal{L}} \varphi(\ulcorner\psi\urcorner) \to \psi$. Since the generalization $\forall x(\varphi(x) \to Tx)$ is supposed to be equivalent to the infinite conjunction, we should be able to infer $\varphi(\ulcorner\psi\urcorner) \to \psi$ from it, for every ψ. That much seems uncontroversial and we therefore propose it as a first condition on the expressibility of infinite conjunctions. While earlier we restricted our attention to infinite conjunctions over truth-free sentences, the following definitions will be more liberal: we allow that the sentences to be generalised may contain the truth predicate themselves.

Definition 4.1.3 (1st condition on expressibility of infinite conjunctions)**.** For all predicates $\varphi(x) \in \mathcal{L}_T$ and sentences $\psi \in \mathcal{L}_T$ the following must hold:

$$\Gamma, \forall x(\varphi(x) \to Tx) \vdash \varphi(\ulcorner\psi\urcorner) \to \psi$$

Assuming that the truth theory and its underlying consequence relation are effective, it is immediately clear why a theory satisfying the above condition is useful in actual reasoning: Under minimal assumptions (namely, transitivity of the consequence relation), the above condition implies that relative to the truth theory every infinite conjunction $\bigwedge_{\psi\in\mathcal{L}_T}(\varphi(\ulcorner\psi\urcorner) \to \psi)$ is reducible or 'finitely axiomatized' by the corresponding generalization $\forall x(\varphi(x) \to Tx)$ in the sense that, whenever a sentence

χ follows from the infinitely many sentences $\varphi(\ulcorner\psi\urcorner) \to \psi$, then χ also follows from the generalization $\forall x(\varphi(x) \to Tx)$.

So far we have looked upon the generalization $\forall x(\varphi(x) \to Tx)$ as expressing the infinite conjunction $\bigwedge_{\psi \in \mathcal{L}_T}(\varphi(\ulcorner\psi\urcorner) \to \psi)$, but it would be equally natural to think of the generalization as expressing the infinite conjunction of the φs, i.e. the conjunction of all sentences ψ such that ψ satisfies the predicate $\varphi(x)$. From this point of view, what the truth predicate should allow us to do is to derive all the φs from the assumption that all φs are true, given an identification of the φs—the generalization $\forall x(\varphi(x) \to Tx)$ ought to 'capture' all the φs. If the deduction theorem (or its semantic counterpart) holds for $\vdash$, our first condition is actually equivalent to this second condition:

Definition 4.1.4 (2nd condition on the expressibility of infinite conjunctions)**.** For all predicates $\varphi(x) \in \mathcal{L}_T$ and sentences $\psi \in \mathcal{L}_T$ we require the following:

$$\Gamma, \forall x(\varphi(x) \to Tx), \varphi(\ulcorner\psi\urcorner) \vdash \psi \tag{TE}$$

It is precisely this inferential property that Horwich has emphasized in the Oscar example. If one wants to assert a sentence or a set of sentences by availing oneself of the truth predicate, then the sentence containing the truth predicate should better imply (relative to the background theory, anyways) all the sentences one initially wanted to assert—otherwise one's assertion fails to serve its purpose. If Jones says that everything that Oscar said is true, we want—given an indentification of what Oscar said—to be able to derive all the statements that Oscar made. Thus Horwich says: "[the generalising] function of truth requires mereley that the generalizations permit us to *derive* the statements to be generalized..." ([46], p. 124)

The assertion 'All φs are true' commits us to all the φs. Now suppose that one of the φs is refutable. Then clearly the claim that all φs are true should be refutable too. For example, if one of the things that Oscar said is '2+2=5', then the assertion that everything that Oscar said is true should be refutable too. Thus, we take it as a third minimal adequacy condition for a truth predicate to express infinite conjunctions that the assertion that all φs are true is refutable, given that one of the φs is refutable and provided that we can identify it as a φ.

Definition 4.1.5 (3rd condition on the expressibility of infinite conjunctions)**.** For all predicates $\varphi(x) \in \mathcal{L}_T$ and sentences $\psi \in \mathcal{L}_T$ we require the following:

$$\Gamma, \varphi(\ulcorner\psi\urcorner), \neg\psi \vdash \neg\forall x(\varphi(x) \to Tx) \tag{TE^C}$$

In a classical context, the third requirement on the generalising function of truth is equivalent to the second requirement, which, given the deduction theorem, is equivalent to the first. However, there might be (and in fact there are) some non-classical logics where this fails to be the case. Therefore, we demand that all three criteria are satisfied. Finally, we say that Γ has the *elimination property* if and only if Γ satisfies the first three conditions on the expressibility of infinite conjunctions. We hope that the precedeeing discussion makes it clear why a theory of truth having the elimination property is useful for finite reasoners: if the truth theory and the consequence relation are effective then generalizations allow us to finitely axiomatize every definable set of sentences by subsuming the predicate defining that set under the truth predicate.

4.1.2. Some observations

Before we have a look at the introduction rules for infinite conjunctions, let us first investigate what features a truth theory needs to posses in order to have the elimination property. For that purpose it is convenient to split the T-schema resp. the Intersubstitutivity Principle into two halfs:

$$T\ulcorner\psi\urcorner \vdash \psi \tag{T-Elim}$$

$$T\ulcorner\psi\urcorner \rightarrow \psi \tag{T-Out}$$

$$\psi \vdash T\ulcorner\psi\urcorner \tag{T-Intro}$$

$$\psi \rightarrow T\ulcorner\psi\urcorner \tag{T-In}$$

The first observation is not very surprising:

Observation 4.1.6. *Let $\Gamma \subseteq \mathcal{L}_T$ be some classical theory of truth where T-Out (or, equivalently, T-Elim) holds. Then Γ has the elimination property.*

Therefore, in classical contexts, T-Out or, equivalently, T-Elim, is *sufficient* for the elimination property. Actually, it is not hard to see that under minimal conditions T-Elim—and, if conditional proof holds, T-Out too—is also a *necessary condition.*

Observation 4.1.7. *Let $\Gamma \subseteq \mathcal{L}_T$ be a theory of truth where identity behaves classically,[1] conditional expressions are true if both antecedent and consequent are true or the former gets a non-designated value, and universal statements are true when all instances are true. If Γ has the elimination property, then T-Elim holds in Γ. Moreover, if conditional proofs hold, then T-Out holds too.*

The conditions imposed on Γ are satisfied in classical logic and in many non-classical ones (the only logic we are aware of that does not satisfy these conditions is Weak Kleene logic, where one of the requirements on the conditional is violated).

[1] This is, the inference from $s = t$ and $\varphi(s)$ to $\varphi(t)$ holds for every formula $\varphi(x)$ and pair of terms s, t.

Proof. Let $\mathcal{M}$ be a model of Γ and let $T\ulcorner\psi\urcorner$ be true in $\mathcal{M}$. Given the conditions imposed on identities, conditional and universal expressions, we know that both $\ulcorner\psi\urcorner = \ulcorner\psi\urcorner$ and $\forall x(x = \ulcorner\psi\urcorner \rightarrow Tx)$ are true in $\mathcal{M}$. Thus, by (TE), we have that ψ is true in $\mathcal{M}$ too. □

What is a bit surprising, however, is that certain non-classical truth theories—in fact, transparent truth theories—do *not* have the elimination property. In order to see how they might fail, we briefly point out what classical inference rules are needed beyond T-Out/T-Elim. The following shows how to derive (TE) in a classical T-Out theory.

1.	$\forall x(\varphi(x) \rightarrow Tx)$	premise 1
2.	$\varphi(\ulcorner\psi\urcorner)$	premise 2
3.	$\varphi(\ulcorner\psi\urcorner) \rightarrow T\ulcorner\psi\urcorner$	1, ∀-E
4.	$T\ulcorner\psi\urcorner$	2, 3, Modus Ponens
5.	$T\ulcorner\psi\urcorner \rightarrow \psi$	(T-Out)
6.	ψ	4, 5, Modus Ponens

(TE^C) can be derived as follows:

1.	$\varphi(\ulcorner\psi\urcorner)$	premise 1
2.	$\neg\psi$	premise 2
3.	$T\ulcorner\psi\urcorner \rightarrow \psi$	(T-Out)
4.	$\neg T\ulcorner\psi\urcorner$	2, 3, Modus Tollens
5.	$\varphi(\ulcorner\psi\urcorner) \wedge \neg T\ulcorner\psi\urcorner)$	2, 4, ∧-I
6.	$\neg(\varphi(\ulcorner\psi\urcorner) \rightarrow \neg T\ulcorner\psi\urcorner)$	5, De Morgan
7.	$\exists x\neg(\varphi(x) \rightarrow Tx)$	6, ∃-I
8.	$\neg\forall x(\varphi(x) \rightarrow Tx)$	7, quantor negation

One does not have to look far to find examples of non-classical theories of disquotational truth in which some of those inferences are invalid. As a consequence, such truth theories actually do *not* have the elimination property. Since weakening of classical logic is not something that is done lightly, the following observations cast severe doubt on the adequacy of these non-classical logics. For what is the point of weakening classical logic if the resulting truth theory does not enable the truth predicate to serve its purpose?

As a first example, let us consider Kripke's fixed-point theory with the Weak Kleene scheme. Although it satisfies the Intersubstitutivity Principle, this theory does not allow us to infer $\exists x\chi$ from $\chi(t)$ unless all instances of $\chi(x)$ have a definite truth value. (Under the Weak Kleene scheme, conjunction and universal quantifier take the minmum of their arguments, with the ordering $\frac{1}{2} < 0 < 1$.) Thus, the step from 6 to 7 in the derivation of (TE^C) is invalidated in that theory. One can show that there is no way of repairing the proof:

Proposition 4.1.8. *Kripke's fixed-point theory with the Weak Kleene scheme does not have the elimination property.*

Proof. Let ψ be $0 = 1$ and $\varphi(x)$ be the predicate $x = \ulcorner 0 = 1 \urcorner$. The valuation V_{WK} for any model of the family assigns 0 to the sentence $0 = 1$, thus $V_{WK}(\neg 0 = 1)$=1. By definition of φ, $V_{WK}(\varphi(\ulcorner 0 = 1 \urcorner)) = 1$ as well. Therefore, the premises of (TE^C) are true in every fixed-point model. However, that's not so in the case of the conclusion. To see this, note that one of the instances of the universal statement $\forall x(\varphi(x) \to Tx)$ is the conditional $\varphi(l) \to Tl$ (where $l = \ulcorner T\underset{\cdot}{\neg}l \urcorner$), whose truth value is given by

$$1 - min\{1 - V_{WK}(\varphi(l)), V_{WK}(Tl)\}$$

which is $\frac{1}{2}$, since $V_{WK}(Tl)$ is $\frac{1}{2}$ (recall that under V_{WK}, $\frac{1}{2} < 0 < 1$). Since the truth value of a general statement is the minimum of the values of its instances, $\forall x(\varphi(x) \to Tx)$ gets value $\frac{1}{2}$, and so does $\neg\forall x(\varphi(x) \to Tx)$. □

As another example, it is well-known that Priest's LP (the acronym stands for 'logic of paradox') does not satisfy Modus Ponens. LP is a paraconsistent logic in which Explosion (and therefore Modus Ponens and Disjunctive Syllogism) does not hold. Thus, the steps from 3 to 4 and from 5 to 6 in the derivation of (TE) are not valid in that logic. The logic of LP is based on the Strong Kleene scheme V_{SK}, but with designated values not just 1 but also $\frac{1}{2}$. Thus, in LP $\frac{1}{2}$ is to be understood as *both true and false.* An argument is valid in LP if it preserves designated values.

Proposition 4.1.9. *Let* Γ *be the theory consisting of the T-schema and the Intersubstitutivity Principle formulated over LP. Then* Γ *does not have the elimination prrperty.*

Proof. As is well known, by a fixed-point construction we can show there is a model $(\mathbb{N}, S)$ of Γ such that $V_{LP}(T\ulcorner\chi\urcorner) = V_{LP}(\chi)$ for every sentence χ.[2] Let ψ be $\bot$—a formula that always gets value 0—and $\varphi(x)$ be the predicate $T(x \underset{\cdot}{\leftrightarrow} l)$, where again $l = \ulcorner T\underset{\cdot}{\neg}l\urcorner$. We first show that $\varphi(\ulcorner\chi\urcorner)$ is true in $(\mathbb{N}, S)$ for every χ.

$l = \ulcorner T\underset{\cdot}{\neg}l\urcorner$ and the clause for negation imply that in $(\mathbb{N}, S)$ $V_{LP}(T\underset{\cdot}{\neg}l) = \frac{1}{2}$. Thus, for any formula χ, by the clause for the conditional it follows that $V_{LP}(\chi \to T\underset{\cdot}{\neg}l)$, $V_{LP}(T\underset{\cdot}{\neg}l \to \chi) \geqslant \frac{1}{2}$, which means that $V_{LP}(\chi \leftrightarrow T\underset{\cdot}{\neg}l) \geqslant \frac{1}{2}$ too, by the semantic clause for conjunction. Since $(\mathbb{N}, S)$ is a model of transparent truth, $T(\ulcorner\chi \leftrightarrow T\underset{\cdot}{\neg}l\urcorner)$ also gets value 1 or $\frac{1}{2}$ for every χ, which by the identity between $\ulcorner\chi \leftrightarrow T\underset{\cdot}{\neg}l\urcorner$ and $\ulcorner\chi\urcorner \underset{\cdot}{\leftrightarrow} l$ means that $T(\ulcorner\chi\urcorner \underset{\cdot}{\leftrightarrow} l)$ is true in $(\mathbb{N}, S)$ for every χ. In particular, $V_{LP}(T(\ulcorner\bot\urcorner \underset{\cdot}{\leftrightarrow} l)) = \frac{1}{2}$.

Moreover, $\forall x(T(x \underset{\cdot}{\leftrightarrow} l) \to Tx)$ is true in $(\mathbb{N}, S)$ as well. Since $V_{LP}(T(\ulcorner\chi\urcorner \underset{\cdot}{\leftrightarrow} l)) \geqslant \frac{1}{2}$ for every χ, the clause for the conditional implies that $V_{LP}(T(\ulcorner\chi\urcorner \underset{\cdot}{\leftrightarrow} l) \to \chi) \geqslant \frac{1}{2}$ and,

[2] These are the Kripke fixed-point models based on the Strong Kleene valuation. See Kremer [52] for a general presentation and Beall [6] for a more specific one.

by transparency, $V_{LP}(T(\ulcorner\chi\urcorner \dot{\leftrightarrow} l) \rightarrow T\ulcorner\chi\urcorner) \geqslant \frac{1}{2}$. Thus, the clause for the universal quantifier gives us that $V_{LP}(\forall x(T(x \dot{\leftrightarrow} l) \rightarrow Tx)) = \frac{1}{2}$.

Therefore, both premises—$\forall x(T(x \dot{\leftrightarrow} l) \rightarrow Tx)$ and $T(\ulcorner\perp\urcorner \dot{\leftrightarrow} l)$—of condition ($T$E) hold in $(\mathbb{N}, S)$ while the conclusion, namely, $\perp$, doesn't. □

The failure of Modus Ponens in LP pushed many paraconsistent-minded philosophers (e.g. Priest [68], Beall [6]) to focus on the search for a 'suitable conditional', i.e. a conditional-like connective that could be added to LP, satisfying not only Modus Ponens but also other *prima facie* desirable principles. The task is far from being trivial, due to Curry paradoxes. Priest [68] adopts a non-contraposible conditional with which he formulates the T-schema. We will call this theory PTT, for 'Priest's Truth Theory'. Since the new conditional satisfies Modus Ponens, Modus Tollens does no longer hold and condition (TE^C) isn't satisfied.

PTT can be somehow extracted from [68]. There, Priest fully endorses the T-schema and argues that it must hold without restriction for the sake of the generalizing function of truth (cf. Priest [68, chap. 4]). Although he works over LP, in order to avoid the problems stated in proposition 4.1.9—viz. the failure of Modus Ponens—he supplements the logic with a new, non-extensional conditional with which he formulates his version of the T-schema. Furthermore, this new conditional, he argues, must be non-contraposible, i.e. $\varphi \rightarrow \psi$ does not necessarily imply $\neg\psi \rightarrow \neg\varphi$. In his own words, "There seems to be no reason why, in general, if φ is a dialetheia [both φ and $\neg\varphi$ are true in a model], $T\ulcorner\varphi\urcorner$ is too. If φ is a dialetheia, $T\ulcorner\varphi\urcorner$ is certainly true, but it might be simply true, and not also false" (Priest [68, p. 79]). As a consequence, PTT won't be a theory of transparent truth.

Let us call Priest's logic 'PL'. Its standard models are quadruples of the form $(\mathbb{N}, S, W, R)$, where W is a set of possible worlds and R a binary relation over W. We assume that each world in W is related to another one by R, R is surjective. PL's valuation scheme V_{PL} behaves exactly like V_{LP} for the extensional connectives, now relativized to a world w. The conditional, instead, is defined in the following way:

- $V^{w}_{PL}(\varphi \rightarrow \psi) \geqslant \frac{1}{2}$ iff, for all $w'Rw$, if $V^{w'}_{PL}(\varphi) \geqslant \frac{1}{2}$, then $V^{w'}_{PL}(\psi) \geqslant \frac{1}{2}$

This clause leaves a lot of room for falsifying conditionals. A sentence is true in a model if and only if it gets value equal or greater than $\frac{1}{2}$ in every world of the model, and logical consequence is defined as in LP.

Let PTT be PA formulated over PL plus the T-schema.

Proposition 4.1.10. PTT *does not have the elimination property.*

Proof. Let $\varphi(x)$ be $x = l$ and ψ be $T\dot{\neg}l$. Since the identity statement $\ulcorner T\dot{\neg}l\urcorner = l$ is true in every model and, by the T-schema and the clauses for negation and the conditional, both $T\dot{\neg}l$ and $\neg T\dot{\neg}l$ get value $\frac{1}{2}$. Thus, both premises of condition ($T\mathrm{E}^C$) are true in every model. We now show that the conclusion—$\neg\forall x(x = l \to Tx)$—fails to be true in some models.

Given any standard model of PTT, the T-schema implies that at every world w, $v^w_{PL}(Tl \leftrightarrow T\dot{\neg}l) \geqslant \frac{1}{2}$, which together with the semantic clause for the conditional in turn gives us that, at every world $w'Rw$, $V^{w'}_{PL}(Tl) \geqslant \frac{1}{2}$ iff $V^{w'}_{PL}(T\dot{\neg}l) \geqslant \frac{1}{2}$. Since the latter holds at every world, the former must hold too, and so Tl must get either value 1 or $\frac{1}{2}$ in every world. But then, for every term t and every world w, $V^w_{PL}(t = l \to Tt) \geqslant \frac{1}{2}$ too, for if at some world $w'Rw$ $V^{w'}_{PL}(t = l) \geqslant \frac{1}{2}$, then $t = l$ (recall we are working with a standard model), which means that Tt is true at every world, since Tl is so.

Let $(\mathbb{N}, S, W, R)$ be a model of PTT such that, at every world w, $V^w_{PL}(t = l \to Tt) = 1$ for every term t. Then, by the clause for the universal quantifier, $V^w_{PL}(\forall x(x = l \to Tx)) = 1$, which means that $V^w_{PL}(\neg\forall x(x = l \to Tx)) = 0$ at every $w \in W$, by the clause for negation. Therefore, the conclusion of condition ($T\mathrm{E}^C$) is false in the model; $(\mathbb{N}, S, W, R)$ is a counter-model for ($T\mathrm{E}^C$). □

Finally, we cast some doubts on the capacity of the truth predicate of Ripley's STTT (Stric-Tolerant Transparent Truth) to serve its expressive purpose.

One of our elimination conditions is the derivability of ψ from the premises that all φs are true and that ψ is a φ. Clearly, we wish to be able to derive ψ not only from the *hypothesis* that all φs are true but also—and even more so—if we *categorically assert* that all φs are true (for example, if we add such an assertion as an axiom to our overall truth theory or if that claim happens to be a theorem of our theory). That is, we not only demand (TE) to hold but we would also like to be able to derive $\Gamma \vdash \psi$ from $\Gamma \vdash \forall x(\varphi(x) \to Tx)$ and $\Gamma \vdash \varphi(\ulcorner\psi\urcorner)$ (and the same for ($T\mathrm{E}^C$)).

In classical theories and in most non-classical ones both requirements are equivalent. However, a new kind of non-classical truth theories has been under the spotlight lately, viz. those that instead of dropping principles governing logical operators choose to abandon structural rules that shape the very notion of logical consequence (cf. Paoli [63]). One of the main lines of investigation is given by the rejection of the transitivity of the consequence relation. In natural deduction calculi, transitivity is automatically given by the fact that we can put any two proofs together to form a new one. In sequent calculi, however, transitivity might fail, depending on the other rules that are available.

Such is the case of Ripley's transparent theory of truth STTT—which contains both T-Intro and T-Elim—and its underlying logic ST (Strict-Tolerant). The logic ST is given by the valuation scheme V_{LP} of LP (now called V_{ST}), and the truth values keep their meaning. The consequence relation, however, is defined in a different manner: an inference from the members of a set $\Gamma \subseteq \mathcal{L}_T$ to φ is valid ($\Gamma \vDash_{ST} \varphi$) if and only if, in every model where all formulae in Γ have value 1, φ gets value 1 or $\frac{1}{2}$. STTT consists of PA plus the Intersubstitutivity Principle formulated over ST. We will show that, though

$$\Gamma, \forall x(\varphi(x) \rightarrow T(x), \varphi(\ulcorner\psi\urcorner) \;\Rightarrow\; \psi$$

is satisfied in STTT,[3] the failure of transitivity of the consequence relation yields some formulae $\varphi(x)$ and ψ such that

$$\Gamma \;\Rightarrow\; \forall x(\varphi(x) \rightarrow Tx)$$

and

$$\Gamma \;\Rightarrow\; \varphi(\ulcorner\psi\urcorner)$$

but it isn't the case that

$$\Gamma \;\Rightarrow\; \psi$$

Analogously, $(T\mathrm{E}^C)$ fails to hold too, i.e. it only holds when premises are hypothetically but not categorically asserted.

Proposition 4.1.11. *The truth predicate of* STTT *doesn't satisfy conditions (TE) and (TE^C) when premises are categorically asserted.*

Proof. Let $\varphi(x)$ and ψ be as in the proof of proposition 4.1.9. Assume we categorically assert $\forall x(T(x\underset{\cdot}{\leftrightarrow}l) \rightarrow Tx)$ and $T(\ulcorner\bot \leftrightarrow T\underset{\cdot}{\neg}l\urcorner)$, i.e., we reason in a theory Th that extends STTT with these two axioms. Then, $\vDash_{Th} \forall x(T(x\underset{\cdot}{\leftrightarrow}l) \rightarrow Tx)$ and $\vDash_{Th} T(\ulcorner\bot \leftrightarrow T\underset{\cdot}{\neg}l\urcorner)$, but $\nvDash_{Th} \bot$, the resulting system isn't trivial.

To see this, let $(\mathbb{N}, S)$ be a model of STTT. As for LP theories of transparent truth, $V_{ST}(\forall x(T(x\underset{\cdot}{\leftrightarrow}l) \rightarrow Tx)) = V_{ST}(T(\ulcorner\bot \leftrightarrow T\underset{\cdot}{\neg}l\urcorner)) = \frac{1}{2}$ in every model of STTT. Thus, every model of STTT is also a model of Th. But not every model of STTT is the trivial model (where all formulae are true). Thus, at least one of these models is such that $V_{ST}(\bot) = 0$. The proof of the failure of condition $(T\mathrm{E}^C)$ for categorically asserted premises is analogous. □

We have now seen that neither the Intersubstitutivity Principle nor the unrestricted T-schema are by themselves sufficient for the truth predicate to serve its purpose. There are non-classical truth theories that enjoy a disquotational truth predicate that, none the less, do not have the elimination property. In some cases, this flaw can be overcome by developing a decent conditional that is not defined in terms of negation and disjunction. However, observation 4.1.6 indicates that there might be no reason to embrace a non-classical logic in the first place.

[3] $\Rightarrow$ stands for the sequent arrow here.

4.2. More on infinite conjunctions

Let us now discuss if, in addition to the elimination property, we can also define a sensible introduction property, this time corresponding to the introduction rules for infinite conjunctions. Then we could say that a truth theory enables us to express infinite conjunctions if and only if it has both the elimination and the introduction property. Given the introduction rule for infinite conjunctions, we can infer $\bigwedge_{\psi\in\mathcal{L}}\varphi(\ulcorner\psi\urcorner)\to\psi$ from $\{\varphi(\ulcorner\psi\urcorner)\to\psi|\psi\in\mathcal{L}\}$. Obviously, we cannot expect that a corresponding rule holds for a truth theory unless we allow some infinitary rule (or, in case we are dealing with a semantic theory of truth, that the theory is already closed under some infinitary rule).

Definition 4.2.1 (4th condition on the expressibility of infinite conjunctions)**.** For all predicates $\varphi(x)\in\mathcal{L}_T$ we require the following:

$$\Gamma,\{\varphi(\ulcorner\psi\urcorner)\to\psi|\psi\in\mathcal{L}_T\}\vdash_\omega\forall x(\varphi(x)\to Tx) \tag{TI}$$

There is a similar, but weaker rule that we can define without invoking the ω-rule. Recall that the 2nd condition on the expressibility of infinite conjunctions requires that the generalization $\forall x(\varphi(x)\to Tx)$ captures all the φs, in the sense that whenever we assume that ψ is a φ, then ψ must be derivable from the generalization (relative to the truth theory). Thus, if Jones says 'Everything that Einstein said is true' then, given an identification of what Einstein said, Jones statement must imply everything that Einstein said. Conversely, assume that what Einstein said was exactly $A_1,\dots,A_n$ and assume furthermore that $A_1,\dots,A_n$ hold indeed. Then we might expect that our truth theory allows us to derive that everything that Einstein said is true. Of course, this only works if the predicate φ applies only to finitely many sentences.

Definition 4.2.2 (5th condition on the expressibility of infinite conjunctions)**.** For all predicates $\varphi(x)\in\mathcal{L}_T$ and all sentences $A_1,\dots,A_n\in\mathcal{L}_T$ we require the following:

$$\Gamma,A_1,\dots,A_n,\forall x(\varphi(x)\leftrightarrow\bigvee_{i\leqslant n}x=\ulcorner A_i\urcorner)\vdash\forall x(\varphi(x)\to Tx), \tag{TI$\restriction$}$$

Let us say that Γ has the *introduction property* if and only if it satisfies the fourth and the fifth condition on the expressibility of infinite conjunctions. And let us say that Γ enables us to express infinite conjunctions, or satisfies the full equivalence between generalizations and infinite conjunctions, if and only if Γ has both the elimination and the introduction property.

The following is easily seen:

Observation 4.2.3. *In classical logic, T-In (or equivalently, T-Intro) is necessary and sufficent for the fourth and fifth requirement on the expressibility of infinite conjunctions.*

Since T-Out and T-In taken together are inconsistent within classical logic, there is no classical truth theory that satisfies the full equivalence between generalizations and infinite conjunctions. Does this spell doom for classical truth theories?

Our answer to the last question is a clear 'No'. In a nutshell, our argument is that all the features that make generalizations useful in actual reasoning are accounted for by the elimination property. In other words: T-Out acconts for all the uses that make the truth predicate indispensable.

A theory that has the elimination property enables us to finitely axiomatize infinite sets of premises by a single expression. What would be the advantage of having a theory that has, in addition, the introduction property? Horwich, in adressing the problem of the paradoxes, points to the following:

> [T]he need to restrict instantiation of the [T-schema] is somewhat in tension with the minimalist thesis about the function of our concept of truth—namely that it enables us to capture schematic generalizations. For, in so far as 'p' is not invariably equivalent to '$\langle p \rangle$ is true', then a generalization of the form 'Every instance of schema S is true' will not invariably entail every instance of S; nor will it always be justified or explained on the basis of those sentences. [...] However, such problematic cases are few and far between; so the utility of truth as a device of generalization is not substantially impaired by their existence.([46, p. 42, fn 21])

Horwich mentions two reasons why the T-schema is needed: first, given the T-biconditionals the generalization 'All φs are true' *entails* every member of φ; secondly, given the T-biconditionals, we can *justify* or *explain* the generalization on the basis of the φs. The first point corresponds to our elimination property; and we have seen that for that purpose T-Out suffices. Horwich's second point corresponds to our introduction property; and it is here where we need T-In. Given T-In, we can justify or explain the generalization $\forall x(\varphi(x) \rightarrow Tx)$ on the basis of the infinitely many premises $\{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi | \psi \in \mathcal{L}_T\}$.

Let us first deal with the case that $\varphi(x)$ applies only to finitely many sentences. For example, assume that the pope said exactly $A_1, \ldots, A_n$ and assume that $A_1, \ldots, A_n$ does indeed hold. Now if our truth theory satisfies the fifth condition on the expressibility of infinite conjunctions, i.e. if it contains T-In, we can conclude that everything the pope said is true.

First, we would like to point out that if generalizations are employed to replace *finite* conjunctions then they are in principle *dispensable*—we can equally use the finite conjunction itself. The main use of a generalization involving the truth predicate is to 'finitely axiomatize' an infinite set. Of course, it is true that in practice we sometimes find ourselves in circumstances where the use of finite generalizations is of value—be it that we are too lazy to repeat the finitely many sentences or that we have problems remembering them. However, one might wonder whether *this* gives us strong reasons for weakening classical logic.

That being said, classical T-Out theorists are of course free to adopt instances of T-In for quite a large number of sentences. (Many parts of this thesis will be concerned with the question of how far we can push the T-schema in classical logic.) If $\varphi(x)$ applies only to such sentences, the generalization can be dervied in an appropriate classical truth theory. Thus, as Horwich said, "such problematic cases are few and far between; so the utility of truth as a device of generalization is not substantially impaired by their existence." Moreover, if $\varphi(x)$ applies to a paradoxical sentence, one may in fact doubt whether the generalization would be justified or ought to be justifiable. To repeat our earlier example, suppose that $\varphi(x)$ applies (amongst others) to the liar sentence. Are we really justified to say that all φs are true eventhough we know that one of them is equivalent to its own untruth?

But maybe there are other cases where we would like to have the equivalence of a generalization and a finite conjunction. Consider the following scenario by Field [26, p. 210]. Suppose you do not remember exactly what Jones said, but you believe that it entails a certain proposition B. Thus, you might say

$$\forall x(\varphi(x) \rightarrow Tx) \rightarrow B, \tag{4.4}$$

where $\varphi(x)$ applies exactly to the sentences uttered by Jones. Then, relative to the assumption that what Jones said is exactly $A_1, \ldots, A_n$, we want the above to imply that

$$A_1 \wedge \ldots \wedge A_n \rightarrow B. \tag{4.5}$$

Field uses this example as an argument against classical truth predicates. He points out that, in order to derive (4.5) from (4.4), A_i and "A_i' is true' need to be intersubstitutable, which won't be the case in general on *any* consistent classical truth theory.

Again, we would first like to point out that in such cases, the use of the truth predicate is in principle dispensable and therefore does not justify weakening classical logic. Secondly, we would like to point out the classical T-Out theorist has some means to deal with Field's problem. It is true that in general (4.4) won't get us to (4.5) unless we have the T-In instances for the A_is at our disposal. But there is *no need* to use (4.4) as a way of expressing (4.5). The latter can be captured by a simple generalization of the form

$$\forall x(\psi(x) \rightarrow Tx) \tag{4.6}$$

Namely, let $\psi(x)$ express that x is the unique sentence obtained by concatenating the conjunction of the φs with the expression '$\rightarrow B$'.[4] Then, in any classical T-Out

[4] More precisely, let $\psi(x)$ be the formula 'for all (finite) sequences of sentences y and all sentences z, w, if (z is a member of y iff $\varphi(z)$) and w is the conjunction of the members of y, then $x = (w \underset{\cdot}{\rightarrow} \ulcorner B \urcorner)$.'

theory, (4.5) is derivable from (4.6) and the assumption that φ applies exactly to $A_1, \ldots, A_n$. This strategy can be generalised to cases of arbitrary complexity.

Our suggestion might seem slightly *ad hoc*, but we rather think of it as a proposal for some new form of regimentation. At any rate, we think a little adhocness outweights the costs of meddling with classical logic.

The above strategy extends to examples of a rather different character. Take, for instance, a definition of knowledge. Epistemologists usually turn to the truth predicate to define knowledge in a non-schematic way. An agent is said to know a sentence just in case she believes it, she is justified in doing so, and, moreover, the sentence is true (and some Gettier condition is satisfied). Formally, epistemologists assert

$$\forall x(K(a,x) \leftrightarrow C(a,x) \wedge Tx) \tag{4.7}$$

instead of the infinitely many instances of the following schema

$$K(a, \ulcorner A \urcorner) \leftrightarrow C(a, \ulcorner A \urcorner) \wedge A \tag{4.8}$$

where $C(x,a)$ resumes all conditions for knowledge except truth.

Suppose now there is an agent a and a sentence A such that $C(a, \ulcorner A \urcorner) \wedge A$. We would like to be able to conclude that a knows A, but without the corresponding instance of T-In, (4.7) does not get us there. It seems that a disquotational truth predicate is required. However, as before, there is no need to generalise on the instances of (4.8) by (4.7). We may well do so by a generalization of the form

$$\forall x(\varphi(a,x) \rightarrow Tx) \tag{4.9}$$

where $\varphi(a,x)$ is true exactly of all instances of (4.8). As expected, any classical T-Out theory will allow us to infer that a knows A from (4.9).

We concede that the above 'definition' of knowledge has it shortcomings. The predicate K is no longer eliminable, and the definition does not satisfy the condition of being non-creative. But again, one should weight the costs of this against introducing a non-classical truth predicate into the definition of knowledge. The non-classicality of truth is contagious: its non-classicality would spread out and make knowledge a non-classical predicate too.

Let us now deal with the justification of generalizations $\forall x(\varphi(x) \rightarrow Tx)$, where $\varphi(x)$ applies to infinitely many things. Obviously, 'justification' here cannot mean 'provability'. Horwich writes ([47, p. 84, fn 14]):

> As for the minimalist, he needs to show how general facts about truth could be explained in terms of what he alleges to be the *basic* facts about truth—i.e. facts of the form, $[T\ulcorner\varphi\urcorner \leftrightarrow \varphi]$. But he is licensed to cite further explanatory factors (as long as they do not concern truth). And this license yields a solution. For it is possible to suppose that there is a truth-preserving rule of inference that will

> take us from a set of premises attributing to each proposition of a certain form some property, G, to the conclusion that *all* propositions have property G. And this rule—not *logically* valid, but none the less necessarily truth-preserving given the nature of propositions—enables the general facts about truth to be explained by their instances. [...] The idea comes from Tarski himself that generalizations about truth may be deduced from their instances by means of some such rule ("infinite induction").

Thus, the idea is that we can justify a generalization by deriving it with the help of some ω-rule. Assume, for example, that we want to justify a generalization such as 'All sentences of the form $A \to A$ are true'. Presumably, our base theory already proves all sentences of the form $A \to A$. Thus, given T-In, we get $T\ulcorner A \to A\urcorner$ and, assuming the ω-rule, we can derive the generalization 'All sentences of the form $A \to A$ are true'. But obviously, we never use the ω-rule. So why the trouble of first going through T-In in order to justify the generalization? Is it not enough to say: "The generalization 'All sentences of the form $A \to A$ are true' captures or finitely axiomatizes the infinite set $\{A \to A | A \in \mathcal{L}_T\}$. We are justified in believing the latter, so we are justified in adopting the generalization."?

4.3. Reflecting on classical truth

In the previous sections we have established T-Out as an attractive—in fact, a necessary—principle for classical truth theories. Notice, however, that any T-Out theory decides the liar.

Proposition 4.3.1. PA + *T-Out* $\vdash \lambda$.

Proof. Let λ be such that $\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$.

1.	$T\ulcorner\lambda\urcorner \to \lambda$	T-Out
2.	$\neg T\ulcorner\lambda\urcorner \to \lambda$	def. of λ
3.	λ	1, 2, logic

Notice also that PA + T-Out $\vdash \neg T\ulcorner\lambda\urcorner$. □

This means that any T-Out theory will have theorems (such as λ) that the theory itself declares *untrue*. One reason one might think that this poses a problem is that the usual way of expressing agreement with a theory is to say 'All theorems of the theory $\mathcal{S}$ are true'. This is also known as the global reflection principle for $\mathcal{S}$, formally

$$\forall x(Prov_{\mathcal{S}}(x) \to Tx) \qquad (\text{GRP}_{\mathcal{S}})$$

where we assume that $\mathcal{S}$ is a recursively axiomatized theory. Now the problem is that any theory that has untrue consequences will be inconsistent with its own global reflection principle.

Proposition 4.3.2. *Suppose that $\mathcal{S}$ is a r.e. theory extending* PA. *If $\mathcal{S} \vdash \lambda$, then $\mathcal{S} + GRP_{\mathcal{S}}$ is inconsistent.*

Proof. Assume $\mathcal{S} \vdash \lambda$ and therefore, by definition of the liar, also $\mathcal{S} \vdash \neg T\ulcorner\lambda\urcorner$. Since $\mathcal{S}$ is r.e. we have $\mathcal{S} \vdash Prov_{\mathcal{S}}(\ulcorner\lambda\urcorner)$ and by $\mathrm{GRP}_{\mathcal{S}}$ it follows that $\mathcal{S} + GRP_{\mathcal{S}} \vdash T\ulcorner\lambda\urcorner$. Thus $\mathcal{S}$+$\mathrm{GRP}_{\mathcal{S}}$ is inconsistent. □

In particular, we have:

Corollary 4.3.3. *Any theory extending* PA + *T-Out is inconsistent with its own global reflection principle.*

Field [26] notices that T-Out theorists might not only have problems with expressing agreement but also with expressing disagreement.

> [Assume that Jones] puts forward a quite elaborate gap theory involving T-Out. And suppose that I disagree with this theory overall, but can't quite decide which specific claims of the theory are problematic. It is natural for me to express my disagreement by saying 'Not everything in Jones' theory is true'. But this doesn't serve its purpose: since Jones himself, as a gap theorist, believes that important parts of his own theory aren't true, I haven't succeeded in expressing disagreement.
>
> Alternatively, suppose that Jones himself thinks that Brown's theory is wrong, but isn't quite sure which claims of it are wrong. Then he certainly can't express his disagreement by saying 'Not everything in Brown's theory is true', since by his own lights that doesn't differentiate Brown's theory from his own. ([26, p. 140])

I do not find these arguments very compelling. Frankly, there is no reason why we should express our agreement by saying 'Everything in Jones' theory is true' or our disagreement by saying 'Not everything in Brown's theory is true'. If Jones disagrees with Brown's theory, Jones will suspect that Brown's theory is *false*, i.e. that there is some sentence φ such that φ is part of Brown's theory, but $\neg\varphi$. And this can be expressed by saying 'Something in Brown's theory is false'—and *that* claim does differentiate Brown's theory from Jones. Similarily, Jones can express agreement with her own theory by saying 'Nothing in my theory is false' or 'Everything in my theory is non-false', formally

$$\forall x(Prov_{\mathcal{S}}(x) \rightarrow \neg T \dot{\neg} x) \qquad (\mathrm{GRP}^*_{\mathcal{S}})$$

I call this the modified global reflection principle; it states that no theorem of $\mathcal{S}$ is false. The following shows that the modified global reflection principle can be consistently added to any T-Out theory that has a standard model.

Proposition 4.3.4. *If $\mathcal{S} \supseteq \mathsf{PA}$ contains T-Out and has a standard model, then $\mathcal{S} + GRP^*_{\mathcal{S}}$ has a standard model too.*

Proof. Assume otherwise. Then there must be a standard model M such that $M \vDash \mathcal{S}$ and $M \vDash \exists x(Prov_{\mathcal{S}}(x) \wedge T\dot{\neg}x)$. Since M is standard, there must be a φ such that $M \vDash Prov_{\mathcal{S}}(\ulcorner\varphi\urcorner) \wedge T\ulcorner\neg\varphi\urcorner$. By T-Out, $M \vDash \neg\varphi$. But since M is standard, $M \vDash Prov_{\mathcal{S}}(\ulcorner\varphi\urcorner)$ implies $\mathcal{S} \vdash \varphi$ and therefore $M \vDash \varphi$. This contradicts $M \vDash \neg\varphi$. □

One attractive feature of the ordinary global reflection principle is that it implies the consistency of the system in question. The modified GRP does the same job.

Proposition 4.3.5. *If $\mathsf{PA} \subseteq \mathcal{S}$ and $\mathcal{S} \vdash T\ulcorner\neg 0 = 1\urcorner$, then $\mathcal{S} + GRP^*_{\mathcal{S}} \vdash Con(\mathcal{S})$.*

Proof. By universal instantiation, $\mathcal{S} + GRP^*_{\mathcal{S}} \vdash Prov_{\mathcal{S}}(\ulcorner 0 = 1\urcorner) \to \neg T\ulcorner\neg 0 = 1\urcorner$. But since $\mathcal{S} \vdash T\ulcorner\neg 0 = 1\urcorner$, it follows that $\mathcal{S} + GRP^*_{\mathcal{S}} \vdash \neg Prov_{\mathcal{S}}(\ulcorner 0 = 1\urcorner)$. □

Asserting untrue sentences is something that seems to be in conflict with our ordinary norms of assertion and denial—norms that are built around principles like 'Assert only sentences that are true and deny only sentences that are false'. Notice that this is not only a problem for T-Out theorists but a problem that concerns all classical truth theorists. A classical logician is committed to accept:

$$\lambda \vee \neg\lambda$$

Although a classical logician may remain agnostic between both disjuncts, she cannot reject both. A little computation shows that the above disjunction actually implies

$$(\lambda \wedge \neg T\ulcorner\lambda\urcorner) \vee (\neg\lambda \wedge T\ulcorner\lambda\urcorner)$$

But it seems that, if we have committed ourselves to a disjunction, then we should be prepared to embrace one of its disjuncts. Even if we remain agnostic between boths disjuncts, our overall theory should be compatible with at least one of the disjuncts. From a classical point of view, it therefore must be possible either to assert a sentence that is not true or to deny a sentence that is true. Accordingly, classical logic is incompatible with the principle 'Only assert sentences that are true, only deny sentences that are false' (if by denying a sentence we mean asserting its negation).

Asserting that the Liar is not true requires that we change our norms of assertion. Tim Maudlin [57] [58] has argued long since that we can avoid revenge by changing the norms of assertion. We have to reject the principle 'Assert only sentences that are true and deny only sentences that are false' and instead lay down principles that allow us to assert (some) sentences that are not true. Then we can express the defectiveness of the Liar just by saying 'The Liar is neither true nor false'. Of course, we cannot truly say so: saying 'It is true that the Liar is neither true nor false' yields a contradiction. Maudlin proposes the following principles:

- Any true sentence is assertible.
- No false sentence is assertible.
- For any sentence φ: not both φ and $\neg\varphi$ are assertible.
- For any sentence φ: either φ or $\neg\varphi$ is assertible.

The third item states that the rules of assertion and denial are pragmatically coherent, while the fourth item states they are complete. Maudlin speaks of the above set of rules as an ideal, albeit one that can never be achieved. Consider the following sentence:

$$\text{The sentence marked (1) is not assertible} \tag{1}$$

Maudlin argues for the inconsistency of the above rules of assertion as follows. Suppose that (1) is assertible; then the above rules would allow the assertion of a falsehood. Suppose (1) is not assertible; then the rules forbid the assertion of a true sentence. If the rules allow the assertion of both (1) and its negation, or forbid the assertion of either (1) or its negation, then the rules are either pragmatically incoherent or incomplete. Thus the above rules express an ideal that can never be achieved.

However, this reasoning depends on the equivalence of (1) and '(1) is true'. If we stay classical, we can just reject the T-biconditional for (1) and its negation, and thus we can avoid this revenge problem altogether. Classical T-Out is consistent with the above norms of assertion.

A classical truth theorist cannot adopt the normative rule 'Only assert sentences that are true, only deny sentences that are false'. Thus it is inevitable that she embraces rules of the sort described above. Horsten, however, has argued against them as follows:

> The trouble with [...] Maudlin's assertion rules [is that they] are not closely related to any rules of assertion that are proposed in the literature. They do not belong to the usual candidates, such as 'assert only what is true', 'assert only what you know', or 'assert only what you rationally believe'. A proposal of extraordinary rules of assertion such as Maudlin's seems to be badly in need of independent support. Otherwise, this proposal has an air of ad hockery around it. ([45, p. 128])

Frankly, I assume that the rules of assertion usually proposed in the literature differ from those of Maudlin precisely because they do not take self-referential sentences into account. I think it is fair to say that once we take self-referentiality into account, none of our traditional rules will survive. For example, *Moore's principle* (cf. [43]) states that

It is incoherent to assert 'φ, but I don't believe φ'

This principle seems uncontroversial as long as we are dealing with ordinary sentences. How can you honestly assert 'Snow is white' while denying to believe that snow is white? But consider a sentence φ that says of itself that I don't believe it. Now because I sincerely reject φ (being suspicious of self-referential statements), I say so (and in fact, truly say so): 'I don't believe φ'. But then I have asserted a sentence while simultaneously denying that I believe that sentence. I don't see how this would make me an irrational person. There is no problem with violating Moore's principle, simply because its formulation didn't take self-referential sentences into account—sentences that 'diagonalize out' of the principle.

Part II.

Grounded truth

5. A graph-theoretic analysis of the semantic paradoxes

If we want a classical theory of untyped truth, we need to know which axioms we can consistently assume. We have already seen that T-Out should be part of our theory, but T-Out alone is not enough. We want e.g. some instances of T-In (the converse of T-Out), so that we have some T-biconditionals, and hence we need to know which instances of the T-schema we can safely assume, and which are paradoxical. This is not an all-or-nothing affair. There are certain pairs of sentences φ, ψ such that we can consistently assume the T-biconditional either for φ or for ψ but not for both. In fact, for every sentence φ there is a pair ψ_1, ψ_2 such that our theory will be inconsistent if the T-biconditionals for both ψ_1, ψ_2 is assumed. This is a consequence of McGee's trick:

Proposition 5.0.6 (McGee [59]). *For every sentence $\varphi \in \mathcal{L}_T$, there is a sentence ψ_φ such that the* $\mathsf{PA} + T\ulcorner\psi_\varphi\urcorner \leftrightarrow \psi_\varphi$ *proves φ.*

Proof. By the diagonal lemma, there is a sentence χ such that

$$\chi \leftrightarrow (T\ulcorner\chi\urcorner \leftrightarrow \varphi)$$

Propositional logic yields

$$(T\ulcorner\chi\urcorner \leftrightarrow \chi) \leftrightarrow \varphi$$

Let ψ_φ be such a χ. □

Thus, the T-biconditionals for $\psi_\varphi, \psi_{\neg\varphi}$ are jointly inconsistent. If φ is provable in our theory, then the T-biconditional for ψ_φ will automatically be a consequence of our theory too while that for $\psi_{\neg\varphi}$ will be refutable. But if φ is undecidable, we probably ought to be cautious and assume the T-biconditional for neither ψ_φ nor $\psi_{\neg\varphi}$. One has the feeling that all potentially dangerous sentences involve some kind of self-reference or circularity, as obtained by the diagonal lemma or some similar device, such as Kleene's recursion theorem, but it is really hard to pin that idea down in *syntactical* terms. One natural idea would be to say that a sentence is self-referential if and only if it is equivalent to a sentence that contains the Gödelnumber of that sentence within the scope of the truth predicate. According to that definition, both ψ_φ and $\psi_{\neg\varphi}$ are self-referential. This definition, however, is trivialized by the fact that *every* sentence is equivalent to a sentence containing its code in the scope of the truth predicate. For every sentence φ is logically equivalent to $(T\ulcorner\varphi\urcorner \vee \neg T\ulcorner\varphi\urcorner) \wedge \varphi$. In the present chapter, we therefore try to characterize the safe and the potentially paradoxical sentences by semantic (i.e. model-theoretic) means. Our aim is not to give an explanation of 'what is going wrong' in the liar reasoning. Rather, our goal is to demarcate—in terms of semantically defined notions such as self-reference and circularity—a set of sentences for which it is dangerous to assume the corresponding T-biconditionals.

5.1. Reference and paradox

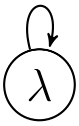
Self-reference is certainly not sufficient for paradox. For example, the T-biconditional for the truth-teller τ with $\mathsf{PA} \vdash \tau \leftrightarrow T\ulcorner\tau\urcorner$ is not only consistent over PA, it is in fact a theorem of PA. But is self-reference necessary for paradox? In 1993, Yablo [96] argued that this is not the case, drawing on the now famous example of an infinite sequence of sentences each of which says that all the sentences appearing later in the sequence are false.

$$\begin{array}{ll} Y(\overline{1}): & \forall x > \overline{1}\ \neg T\ulcorner Y(\dot{x})\urcorner \\ Y(\overline{2}): & \forall x > \overline{2}\ \neg T\ulcorner Y(\dot{x})\urcorner \\ Y(\overline{3}): & \forall x > \overline{3}\ \neg T\ulcorner Y(\dot{x})\urcorner \\ \text{etc.} & \end{array}$$

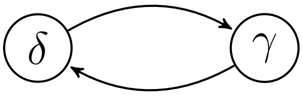
There has been an extensive debate about whether there may be hidden forms of circularity or self-reference in Yablo's paradox or not (Priest [67], Sorensen [87], Beall [3], Cook [15], Picollo [64]). Clearly, formalizing Yablo's paradox in arithmetic usually invokes some fixed point construction, as obtained by Gödel's diagonal lemma or Kleene's recursion theorem. Thus, as some authors have argued (e.g. Priest [67]), the whole sequence is endowed with *some kind* of circularity being inherent to such fixed-point constructions. But there also seems to be a sense of 'circularity' in which the Yablo sentences are clearly not circular—namely, when we think of what the Yablo sentences are *about* (or refer to).

In 1970, about two decades before Yablo's discovery, Hans Herzberger [41] argued that there are referential patterns other than circularity that should be counted as pathological. According to his approach, any sentence has a domain, the *set of objects it is about.* For example, the Liar is about itself; a sentence of the form 'All Φs are true' is about the Φs. Of course, a sentence may contain objects in its domain that are sentences themselves and which are about further sentences etc. Herzberger concedes that 'the general notion of a domain is more readily indicated than explicated'. But let us assume for the moment that we have a method of assigning to each sentence its domain. Say that φ is about ψ iff ψ is in the domain of φ. Let us call a sentence φ *directly self-referential* iff φ is about φ and call a sentence φ *indirectly self-referential* iff φ is about ψ_1 and ψ_1 is about ψ_2,..., and ψ_n is about φ. Finally, call a sentence *circular* if it is either directly or indirectly self-referential.

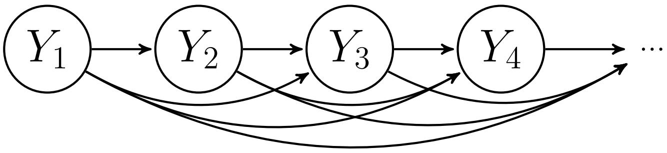
Clearly, our intuition tells us that the liar sentence is about itself while each member of Yabo's sequence is about all the sentences appearing later in the sequence: The liar sentence and all liar cycles are circular while no member of the Yablo sequence is circular. (This judgement depends on a basic intuition we have about the aboutness relation. It can be expressed by the following rule: a sentence of the form $\forall x(\varphi(x) \to \neg T(x))$, where φ does not contain the predicate symbol T, is *not* about ψ unless ψ satisfies the formula $\varphi(x)$.) We can depict our intuitions as follows:



The liar graph



Graph of a liar cycle of arity 2



The Yablo graph

An interesting answer to the question 'Why are some sentences paradoxical while others are not?' thus might look like:

- Some sentences are paradoxical because of their position in the *reference-graph of our language*, i.e. in the directed graph whose vertices are the sentences of the language, two sentences φ and ψ being joined by an arc iff φ bears the relation of reference (aboutness) to ψ.

But which are the paradoxical nodes of the reference-graph? And can they be characterized in graph-theoretic terms? We shall call any approach to semantic paradoxes that is concerned with identifying paradoxical reference patterns a *reference-based theory of semantic paradoxes*. In order to develop a reference-based account we have to

1. give a rigorous definition of the reference (aboutness) relation,
2. give a rigorous definition of paradoxicality, and
3. specify a graph-theoretic property that determines what nodes in the reference-graph are paradoxical.

A good reference-based theory of semantic paradoxes should make us better understand how these concepts are interconnected: The notion of aboutness between sentences; a graph-theoretic property defining a class of critical reference patterns; the notion of a (potentially) paradoxical sentence. Any intuitions we have about any of these notions can shed some light on the others. An explication of any one of them always depends on explicating the others.

There have been several, quite interesting approaches to characterize the notion of a paradoxical sentence for *infinitary propositional* languages by graph-theoretic means (cf. Cook [14], Rabern, Rabern and Macauley [75]). A natural question then is whether such a characterization is also available for first-order languages such as the language of Peano arithmetic. The problem, of course, is that it is far from clear how to define an aboutness relation (equivalently, to explicate the notion of a domain) for arbitrary sentences containing quantifiers. Our proposal is to identify the domain of a sentence of the language $\mathcal{L}_T$ with its dependence set in the sense of Leitgeb [55].

Outline of the remainder of the chapter

In section 5.2 we introduce the basic concepts of Leitgeb's paper on semantic dependence. It will be shown that Leitgeb's theory can be treated within the framework of Kripke's fixed-point semantics. In section 5.3, we show how to define unique reference-graphs (called 'sensitivity-graphs') for those sentences that do possess a canonical dependence set. All of the paradoxes that are usually discussed in the literature—e.g. the liar, liar cycles, Curry's and Yablo's paradox—fall under that category. We also prove some theorems concerning (ω-)consistent subsets of the T-schema in terms of sensitivity.

In section 5.4 we introduce a game-theoretic semantics for Kripke's theory of truth. In section 5.4.1 we define, for any sentence φ and set of sentences S, a grounding game $\mathcal{G}_G(\varphi, S)$ such that φ is grounded in S (i.e. is an element of the fixed point generated by S) if and only if player ($\exists$) has a winning strategy in the game $\mathcal{G}_G(\varphi, S)$. We then show how the strategies available in this game can be used to define an infinite family of reference-graphs for the sentence in question. These reference-graphs can be seen as a generalization of the sensitivity-graphs of section 5.3. In case φ has a least dependence set, there will be a canonical reference-graph among the infinite familiy, and that canonical reference-graph will coincide with the sensitivity-graph of that sentence. We then use our machinery to show that a sentence is grounded if and only if it has a well-founded reference-graph.

In section 5.4.2 we define, for any sentence φ and partial model $\mathcal{F}$, a verification (falsification) game such that φ is true (false) in the fixed-point generated by $\mathcal{F}$ if and only if player $(\exists)$ has a winning strategy in the verification (falsification) game for φ and $\mathcal{F}$.

In section 5.4.3 we apply our machinery to obtain some graph-theoretic descriptions of the Kripke-paradoxical sentences. We show, amongst others, that if a sentence is Kripke-paradoxical, then each of its reference-graphs contains either a directed cycle or infinitely many double paths. We conclude with a conjecture that states, roughly, that every paradox is reducible either to the liar or Yablo's paradox.

5.2. Semantic dependence. Leitgeb (2005)

In [55], Leitgeb aims to give a definition of truth for those sentences that are grounded or, in other words, that depend (directly or indirectly) on non-semantic states of affairs only. What distinguishes his approach from Kripke's is that, first, Leitgeb is more interested in a notion of truth for classical languages and, second, that the notion of groundedness comes first and truth is derived, whereas on Kripke's approach, the order is reversed. The central notion for the construction of the set of grounded sentences is the notion of dependence (or determination):

Definition 5.2.1. φ *depends on* Φ iff for all Ψ_1, Ψ_2: if $\Psi_1 \cap \Phi = \Psi_2 \cap \Phi$, then $(\mathbb{N}, \Psi_1) \vDash \varphi$ iff $(\mathbb{N}, \Psi_2) \vDash \varphi$.

Thus, a sentence φ is determined by a collection of sentences Φ iff for all Ψ_1, Ψ_2: if Ψ_1, Ψ_2 agree on Φ, then they assign the same truth-value to φ. Leitgeb has given the following equivalent definition of dependence.

Proposition 5.2.2 (Leitgeb). *φ depends on Φ iff for all Ψ:*

$$(\mathbb{N}, \Psi) \vDash \varphi \Leftrightarrow (\mathbb{N}, \Psi \cap \Phi) \vDash \varphi$$

Thus, intuitively, a sentence φ depends on a set of sentences Θ iff all the objects that are relevant for the evaluation of φ are among Φ.

Theorem 5.2.3 (Leitgeb). *The operator $D(\Phi) = \{\psi | \psi$ depends on $\Phi\}$ is monotone: If $\Phi \subseteq \Psi$, then $D(\Phi) \subseteq D(\Psi)$.*

Let us call a set Φ *D-sound* iff $\Phi \subseteq D(\Phi)$. The monotonicity of the operator D implies the existence of fixed points. Leitgeb identifies the sentences that depend on non-semantic states of affairs only with the sentences in the minimal fixed point of D.

Definition 5.2.4. We inductively define: $D_0(S) = S$, $D_{\alpha+1}(S) = D(D_\alpha(S))$ and $D_\gamma(S) = \bigcup_{\alpha<\gamma} D_\alpha(S)$ for limit ordinals γ. We call $G(S) := \bigcup_{\alpha\in On} D_\alpha(S)$ the set of sentences that are *grounded in S*. A sentence is called *grounded* (simpliciter) iff it is grounded in the empty set. We write G instead of $G(\varnothing)$ and G_α instead of $D_\alpha(\varnothing)$.

Dependence is monotone: if φ depends on Φ, then φ also depends on any superset $\Psi \supseteq \Phi$. Some sentences have a least dependence set:

Definition 5.2.5. A sentence φ has *essential dependence* iff there is a set Φ such that φ depends on Φ but φ does not depend on any proper subset $\Psi \subset \Phi$. Otherwise we say that φ *lacks* essential dependence.

Examples of sentences with and without essential dependence sets will be given in the next section. The set of grounded sentences has some nice closure properties:

Proposition 5.2.6 (Leitgeb). *1. All arithmetical sentences are grounded.*

2. *All classical tautologies and falsehoods (in the language $\mathcal{L}_T$) are grounded.*

3. *φ is grounded iff $T\ulcorner\varphi\urcorner$ is grounded.*

4. *φ is grounded iff $\neg\varphi$ is grounded.*

5. *If φ, ψ are grounded, then $\varphi \wedge \psi$ is grounded.*

6. *If φ, ψ are grounded, then $\varphi \vee \psi$ is grounded.*

7. *If, for all n, $\varphi(\overline{n})$ is grounded, then $\forall x\varphi$ is grounded.*

8. *The set of grounded sentences is closed under PAT-equivalence.*

In fact, the sentences mentioned under (1) and (2) come in at the first stage of the hierarchy; (3)-(8) provide closure conditions. We will give some concrete examples of (un)grounded sentences in the next section. Now an interesting extension for the truth predicate can be defined as follows:

Definition 5.2.7. By transfinite recursion define Θ_α for $\alpha \in ON$ by:

1. $\Theta_0 = \varnothing$

2. $\Theta_{\alpha+1} = \{\varphi \in G_{\alpha+1} | (\mathbb{N}, \Theta_\alpha) \vDash \varphi\}$

3. $\Theta_\beta = \bigcup_{\alpha<\beta} \Theta_\alpha$, when β is a limit ordinal.

Theorem 5.2.8 (Leitgeb). $\Theta_\alpha \subseteq \Theta_\beta$ *for all* $\alpha < \beta$.

Consequently, there is an α such that $\Theta_\alpha = \Theta_{\alpha+1}$. We denote this fixed point simply by Θ_∞. This fixed point is a model for the T-biconditionals for all grounded sentences:

Corollary 5.2.9 (Leitgeb)**.** *For all $\varphi \in G$: $(\mathbb{N}, \Theta_\infty) \vDash T\ulcorner\varphi\urcorner \leftrightarrow \varphi$.*

In chapter 7 I will show that Θ_∞ is Π^1_1-complete. This was independently established by Welch [94] using a quite different argument. In what follows, I will give a very simple proof (different both from Welch's and our own in chapter 7) that the first level Θ_1 of Leitgeb's truth hierarchy is a Π^1_1-complete set of integers. Stanislav Speranski demonstrated to me that my argument can be generalized to show that for each $\alpha > 0$, G_α and Θ_α are Π^1_1-complete. We need the following preliminary lemma.

Proposition 5.2.10. $\Theta_1 = A := \{\varphi \in \mathcal{L}_T | \forall S \subseteq \omega : (\mathbb{N}, S) \vDash \varphi\}$

Proof. By definition,

$$\Theta_1 = \{\varphi \in \mathcal{L}_T | \varphi \text{ depends on } \varnothing, (\mathbb{N}, \varnothing) \vDash \varphi\}$$

We first show $\Theta_1 \subseteq A$. Let $\varphi \in \Theta_1$. Then

1. φ depends on $\varnothing$
2. $(\mathbb{N}, \varnothing) \vDash \varphi$

By (1) and Proposition 5.2.2, we have for all $S \subseteq \omega$

$$\begin{aligned}(\mathbb{N}, S) \vDash \varphi &\Leftrightarrow (\mathbb{N}, S \cap \varnothing) \vDash \varphi \\ &\Leftrightarrow (\mathbb{N}, \varnothing) \vDash \varphi, \text{ because } S \cap \varnothing = \varnothing.\end{aligned}$$

So by (2) and the above equivalence, $(\mathbb{N}, S) \vDash \varphi$ for all S, so $\varphi \in A$.

Conversely, let $\varphi \in A$. So $(\mathbb{N}, S) \vDash \varphi$ for all S. In particular, $(\mathbb{N}, \varnothing) \vDash \varphi$, so (2) is satisfied. But (1) is also satisfied: since φ has the same truth value under any interpretation, $(\mathbb{N}, S) \vDash \varphi$ iff $(\mathbb{N}, S \cap \varnothing) \vDash \varphi$, so φ depends on the empty set. □

Theorem 5.2.11. *The set Θ_1 is Π^1_1-complete.*

Proof. By the previous proposition, Θ_1 is simply the set of all $\mathcal{L}_T$-sentences that are true under any interpretation of the truth predicate, i.e.

$$\Theta_1 = \{\varphi \in \mathcal{L}_T | \forall S \subseteq \omega : (\mathbb{N}, S) \vDash \varphi\}$$

For any $\varphi \in \mathcal{L}_T$ let $\widehat{\varphi}$ be the result of replacing all occurrences of the predicate T by the second-order variable X. (Observe that $\widehat{\varphi}$ is an *arithmetical* $\mathcal{L}_2$-formula with exactly X free.) Then observe that Θ_1 is recursively isomorphic to the set of all true Π^1_1-sentences, i.e. $\varphi \in \Theta_1$ iff $\forall X \widehat{\varphi}$ is true in the structure $(\mathbb{N}, \wp(\omega))$, where the second-order quantifer ranges over the elements of $\wp(\omega)$. It is well-known that the set of true Π^1_1-sentences is a Π^1_1-complete set. □

Notice that $\Theta_1 = \mathcal{J}^1_{FV}(\varnothing)^+$, so the above proof also establishes that the first level of the minimal Kripke fixed point under the supervaluational scheme is Π^1_1-complete.

Despite the different conceptual intuitions behind both approaches, it is obvious that Leitgeb's theory has a lot in common with Kripke's fixed-point theory. Leitgeb [55] has shown that $\Theta_\infty \subseteq \mathcal{J}^\infty_{FV}(\varnothing)^+$. In fact, the inclusion is proper. For example, the sentence $T\ulcorner 1 = 1\urcorner \vee \lambda$, where λ is the Liar, enters $\mathcal{J}^\infty_L(\varnothing)^+$ at stage 2, but it never enters G, because the dependence set of the disjunction contains the liar. However, since we know that $1 = 1$ is true, we can also determine the truth value of $T\ulcorner 1 = 1\urcorner \vee \lambda$. As another example, we may take the sentence $T\ulcorner 1 \neq 1\urcorner \wedge \lambda$, which is easily seen to be false (since $1 \neq 1$ is false). This sentence is not in G, but it is again a part of $\mathcal{J}^\infty_{FV}(\varnothing)^+$. A variation of the first example shows that Θ_∞ (in contrast to $\mathcal{J}^\infty_{FV}(\varnothing)^+$) invalidates the claim that Modus Ponens preserves truth:

Proposition 5.2.12. $(\mathbb{N}, \Theta_\infty) \nvDash \forall x \forall y (Sent(x \dot{\to} y) \to (T(x \dot{\to} y) \to (Tx \to Ty)))$

Proof. Notice that $(\mathbb{N}, \Theta_\infty) \vDash T\ulcorner T\ulcorner 1 = 1\urcorner \to T\ulcorner 1 = 1\urcorner \vee \lambda\urcorner$ and $(\mathbb{N}, \Theta_\infty) \vDash T\ulcorner T\ulcorner 1 = 1\urcorner\urcorner$ but $(\mathbb{N}, \Theta_\infty) \nvDash T\ulcorner T\ulcorner 1 = 1\urcorner \vee \lambda\urcorner$, because $T\ulcorner 1 = 1\urcorner \vee \lambda$ is not grounded. □

However, by a little modification we can make Θ_∞ equal to $\mathcal{J}^\infty_{FV}(\varnothing)^+$.[1] This can be done by introducing the notion of *conditional* dependence, and by further restricting the quantifiers in the definition to *consistent* supersets.[2]

Definition 5.2.13. We say that φ c-depends$_\Sigma$ on Φ iff for all consistent $\Psi_1, \Psi_2 \supseteq \Sigma$: if $\Psi_1 \cap \Phi = \Psi_2 \cap \Phi$, then $(\mathbb{N}, \Psi_1) \vDash \varphi$ iff $(\mathbb{N}, \Psi_2) \vDash \varphi$.

Definition 5.2.14. We define by simultaneous transfinite recursion:

1. $G'_0 = \varnothing$
2. $G'_{\alpha+1} = \{\varphi | \varphi \text{ c-depends}_{\Theta'_\alpha} \text{ on } G'_\alpha\}$
3. $G'_\beta = \bigcup_{\alpha<\beta} G'_\alpha$, when β is a limit ordinal.
4. $\Theta'_0 = \varnothing$
5. $\Theta'_{\alpha+1} = \{\varphi \in G'_{\alpha+1} | (\mathbb{N}, \Theta'_\alpha) \vDash \varphi\}$
6. $\Theta'_\beta = \bigcup_{\alpha<\beta} \Theta'_\alpha$, when β is a limit ordinal.

Proposition 5.2.15 (Meadows, Bonnay & Vugt). $\Theta'_\alpha = \mathcal{J}^\alpha_{FV}(\varnothing)^+$ *for all* $\alpha \in ON$.

This relationship between Leitgeb's construction and the minimal supervaluational fixed point is interesting (and will play a role in chapter 6), but still leaves us wondering what the exact relation between Leitgeb's and Kripke's theory is.

[1] Cf. Bonnay & Vugt [90] or Meadows [60].

[2] Conditionality takes care of the first example, consistency takes care of the second example.

Definition 5.2.16. Let $S = (S^+, S^-)$ be a consistent partial model. Define the *Leitgeb valuation* scheme by

$$V_L(S)(\varphi) = \begin{cases} 1, & \text{if } \varphi \text{ depends on } S^+ \cup S^- \text{ and } (\mathbb{N}, S^+) \vDash \varphi \\ 0, & \text{if } \varphi \text{ depends on } S^+ \cup S^- \text{ and } (\mathbb{N}, S^+) \nvDash \varphi \\ \frac{1}{2}, & \text{if } \varphi \text{ does not depend on } S^+ \cup S^- \end{cases}$$

It is easily seen that V_L is a monotonic valuation scheme.

Proposition 5.2.17. *For all α, $\Theta_\alpha = \mathcal{J}_L^\alpha(\varnothing)^+$ and $G_\alpha = \mathcal{J}_L^\alpha(\varnothing)^+ \cup \mathcal{J}_L^\alpha(\varnothing)^-$.*

Proof. By simultaneous transfinite induction on α. We only show the first part of the claim, the other one can be proved similarly. For $\alpha = 0$, we have $G_0 = \Theta_0 = \varnothing = \mathcal{J}_L^0(\varnothing)^+ = \mathcal{J}_L^0(\varnothing)^+ \cup \mathcal{J}_L^0(\varnothing)^-$. If α is a limit, apply the induction hypothesis. So let $\alpha = \beta + 1$. Assume as I.H. that $\Theta_\beta = \mathcal{J}_L^\beta(\varnothing)^+$ and that $G_\beta = \mathcal{J}_L^\beta(\varnothing)^+ \cup \mathcal{J}_L^\beta(\varnothing)^-$. Let $\varphi \in \Theta_{\beta+1}$. By defintion of Θ, this means $(\mathbb{N}, \Theta_\beta) \vDash \varphi$ and $\varphi \in G_{\beta+1}$, i.e. φ depends on G_β. So by definition of V_L, $V_L(\mathcal{J}_L^\beta(\varnothing))(\varphi) = 1$, which implies by definition of Kripke jump that $\varphi \in \mathcal{J}_L(\mathcal{J}_L^\beta(\varnothing))^+ = \mathcal{J}_L^{\beta+1}(\varnothing)^+$. The other direction is proved similar. □

Thus, Leitgeb's theory of truth can be treated within the Kripke framework. There is not only a minimal fixed point of $\mathcal{J}_L$, but also (many) maximal fixed points, a largest intrinsic fixed point etc. Moreover, we also have the notion of Kripke-paradoxicality for V_L (cf. section 3.2): a sentence φ is Kripke-paradoxical iff φ does not receive a definite truth value in any fixed point of $\mathcal{J}_V$. Our aim in the next sections is to characterize the Kripke-paradoxical sentences (relative to V_L) in graph-theoretic terms.

5.3. Sensitivity-graphs

In this section we first assign reference-graphs to those sentences of $\mathcal{L}_T$ that have essential dependence. Then we show how information about the sensitivity-graphs of certain sets of sentences provides us with sufficient conditions for the consistency of certain subsets of the T-schema. In the next section, we provide a method to assign reference-graphs to any sentence of $\mathcal{L}_T$. This method will assign infinitely many reference-graphs to each sentence. In case a sentence has essential dependence, it is possible to single out a canonical reference-graph which will be isomorphic to the sensitivity-graph of that sentence.

We will first introduce a relation, called 'sensitivity', that holds between *single* sentences. This notion will provide us with a better grasp on which sentences have essential dependence.

Definition 5.3.1. Let $S \subseteq \omega$ and $\varphi \in \mathcal{L}_T$. Define

$$S_\varphi = \begin{cases} S \setminus \{\varphi\}, \text{if } \varphi \in S \\ S \cup \{\varphi\}, \text{if } \varphi \notin S \end{cases}$$

So S_φ is exactly as S—except for φ. If S contains φ, then S_φ won't, and if S does not contain φ, then S_φ will. In algebraic terms, S_φ is the symmetric difference of S and $\{\varphi\}$, i.e.

$$S_\varphi = (S \cup \{\varphi\}) \setminus (S \cap \{\varphi\})$$

It should be clear that $S = S_{\varphi_\varphi}$, i.e., applying the operation a second time undoes the effect of the first application. In what follows, φ^S denotes the truth-value of φ relative to the model $(\mathbb{N}, S)$, i.e. $\varphi^S = 1$ if $(\mathbb{N}, S) \vDash \varphi$, and $\varphi^S = 0$ otherwise. We also write $Val_S(\varphi)$ instead of φ^S.

Definition 5.3.2. We say that φ is *sensitive* to ψ iff there is an $S \subseteq \omega$ such that:

$$\varphi^S \neq \varphi^{S_\psi}$$

We write $\psi \mathbb{S} \varphi$ if φ is sensitive to ψ. (Notice that we switched the order of the relata!)

We say that φ is *insensitive to* ψ iff φ is not sensitive to ψ, i.e., iff for all $S \subseteq \omega$ we have:

$$\varphi^S = \varphi^{S_\psi}$$

So if φ is insensitive to ψ, then we cannot change the truth-value of φ by adding/removing ψ to/from the extension of the truth-predicate, not matter with which model we start.

Definition 5.3.3. Let $Dom\,(\varphi)$ be the set of sentences to which φ is sensitive, i.e., $Dom\,(\varphi) = \{\psi | \psi \mathbb{S} \varphi\}$.

Whereas the notion of dependence relates sentences and sets of sentences, the notion of sensitivity has only single sentences as relata. Moreover, while the notion of dependence involves a universal quantifier ranging over subsets of ω, the notion of sensitivity only involves an existential quantifier. The fundemantal connection between dependence and sensitivity is given in the following theorem.

Theorem 5.3.4. *Let φ be an $\mathcal{L}_T$-sentence and $\Phi \subseteq \mathcal{L}_T$. Then the following are equivalent:*

1. *φ depends essentially on Φ.*
2. *φ depends on Φ and $\Phi = Dom(\varphi)$.*

A proof of this important result can be found in Beringer & Schindler [7]. Leitgeb calls a sentence φ *self-referential* iff φ is contained in every set on which it depends. Of course, self-reference implies ungroundedness. Checking whether φ is contained in each of its dependence sets can sometimes be laborious; the sensitivity relation provides us with an easier criterion of self-referentiality.

Theorem 5.3.5. *A sentence φ is self-referential iff φ is sensitive to itself.*

Proof. We only show the right-to-left direction. Let φ be sensitive to itself and assume that φ depends$_L$ on Φ. We have to show that $\varphi \in \Phi$. Since φ is sensitive to itself, there is an S such that $\varphi^S \neq \varphi^{S_\varphi}$. Since φ depends$_L$ on Φ, we find that $S \cap \Phi \neq S_\varphi \cap \Phi$. This implies $\varphi \in \Phi$. □

Corollary 5.3.6. *The predicate 'x is self-referential' is Δ^1_1.*

Analogously, we may say that φ is *circular* (or indirectly self-referential) iff there is a ψ (distinct from φ) such that $\psi \mathbb{S} \varphi$ and $\varphi \mathbb{S} \psi$. The following list provides some examples of grounded, ungrounded, self-referential and circular sentences.

Example 5.3.7.

1. *The truth-teller τ with $\mathsf{PA} \vdash \tau \leftrightarrow T\ulcorner\tau\urcorner$ is sensitive to itself, and therefore self-referential and ungrounded.*

   Proof: $Val_\varnothing(\tau) = 0 \neq 1 = Val_{\{\tau\}}(\tau) = Val_{\varnothing_\tau}(\tau)$.

2. *Let A, B be such that $\mathsf{PA} \vdash A \leftrightarrow \neg T\ulcorner B\urcorner, B \leftrightarrow T\ulcorner A\urcorner$. Then A is sensitive to B and B is sensitive to A. Thus both A and B are circular.*

3. *The completeness axiom (Comp) $\forall x\,(Tx \vee T\dot{\neg}x)$ is sensitive to any sentence whatsoever, and therefore ungrounded.*

   Proof: *Let φ be arbitrary. Let $S = \omega \setminus \{\varphi, \neg\varphi\}$. Then $Val_S(Comp) = 0 \neq 1 = Val_{S\cup\{\varphi\}}(Comp) = Val_{S_\varphi}(Comp)$.*

4. *The claim $\varphi \vee \neg\varphi$ is not sensitive to any sentence, and therefore grounded.*

   Proof: *Let φ be arbitrary. Then $Val_S(\varphi) = Val_{S_\psi}(\varphi)$ for all sentences ψ.*

5. *Let the Yablo sequence $\{Y(\overline{n}) | n \in \omega\}$ with $\mathsf{PA} \vdash Y(\overline{n}) \leftrightarrow \forall x > \overline{n} \neg T\ulcorner Y(\dot{x})\urcorner$ be given. Then for all n, $Y(\overline{n})$ is sensitive to all $Y(\overline{m})$ with $m > n$. However, no $Y(\overline{n})$ is circular.*

   Proof: $Val_\varnothing(Y(\overline{n})) = 1 \neq 0 = Val_{\{Y(\overline{m})\}}(Y(\overline{n})) = Val_{\varnothing_{Y(\overline{m})}}(Y(\overline{n}))$.

6. *The sentence $\varphi \doteq (1 = 1 \wedge T\ulcorner\lambda\urcorner)$ is sensitive to λ, where λ is the Liar.*

   Proof: $Val_\varnothing(\varphi) = 0 \neq 1 = Val_{\{\lambda\}}(\varphi) = Val_{\varnothing_\lambda}(\varphi)$.

Now we are in a position to define reference-graphs for those sentences that have essential dependence.

Definition 5.3.8. The *sensitivity-graph* of the language $\mathcal{L}_T$ is the directed graph defined by the sensitivity relation $\mathbb{S}$, and the *sensitivity-graph of a sentence* φ is the smallest (with respect to the subgraph relation) induced subgraph of the sensitivity-graph of $\mathcal{L}_T$ that contains φ and contains with each ψ any sentence χ such that ψ is sensitive to χ (i.e. such that $\chi \mathbb{S} \psi$).[3]

Loosely speaking, the sensitivity-graph of φ is the relation defined by the sensitivity relation restricted to the transitive closure of $\{\varphi\}$ w.r.t. $\mathbb{S}$. It follows from our definitions (and the proofs in the example list given above) that the sensitivity-graph of the liar sentence is isomorphic to the graph shown in picture 1 and that the sensitivity-graph of the Yablo sequence is isomorphic to the graph in picture 3 (at the beginning of this chapter).

Which sets of T-biconditionals can be added to PA without generating a paradox? We state some useful results in terms of sensitivity.

For given S and $\psi_1, \ldots, \psi_n$, let $S_1 = S_{\psi_1}$, $S_{i+1} = (S_i)_{\psi_{i+1}}$. Hence, given S, S_1 results from adding/removing ψ_1 to/from S, S_2 results from adding/removing ψ_2 to/from S_1 and so on. Notice that S_n is identical to the symmetric difference of S and $\{\psi_1, \ldots, \psi_n\}$. The following proposition shows that if φ is insensitive to every member of Φ (where Φ is finite), then the truth-value of φ relative to some extension S remains constant under varying the extension S with respect to members of Φ. The result does not obtain in general if Φ is infinite.

Proposition 5.3.9. *Assume φ is insensitive to $\psi_1, \ldots, \psi_n$. Then for any S, $\varphi^S = \varphi^{S_1} = \ldots = \varphi^{S_n}$.*

Proof. By an easy induction. □

$T \restriction S$ denotes the theory whose axioms are those of PAT plus all T-biconditionals for all members of S.

Proposition 5.3.10. *Let $S \subseteq \mathcal{L}_T$. Assume there is a function $f : S \to \omega$ such that*

$$\text{(*)} \quad \psi \mathbb{S} \varphi \Rightarrow f(\psi) < f(\varphi), \ \text{ for all } \psi, \varphi \in S;$$

then $T \restriction S$ has a model. Moreover, if S is finite, then there is an ω-model.

Proof. By Proposition 5.3.9 and the compactness of first-order logic. So let $A \subseteq S$ be finite. Let $F_n = \{\varphi \in A | f(\varphi) = n\}$. Let $\Gamma_0 = \varnothing$ and $\Gamma_{n+1} = \Gamma_n \cup \{\varphi \in F_n | (\mathbb{N}, \Gamma_n) \vDash \varphi\}$. It follows from (*) and Proposition 5.3.9 that this sequence is monotone. Hence $(\mathbb{N}, \Gamma_{k+1})$ is a model of the T-schema restricted to A, where $k = \max\{n | \exists \varphi \in A.\ f(\varphi) = n\}$. □

[3] See the appendix for definitions of the graph-theoretic notions used in this chapter.

Theorem 5.3.11. *If $S \subseteq \mathcal{L}_T$ contains no $\mathbb{S}$-cycle, then $T \restriction S$ has a model.*

Proof. We will apply the compactness of first-order logic. So let $A \subseteq S$ be finite. Define $f : A \to \omega$ as follows: If $\varphi \in A$ is not sensitive to any $\psi \in A$, then $f(\varphi) = 0$. For any other $\varphi \in A$, let $f(\varphi) = \max\{f(\psi) + 1 | \psi \mathbb{S} \varphi, \psi \in A\}$. Since S contains no $\mathbb{S}$-cycle and A is finite, it follows that f satisfies (*). So the T-schema restricted to A has a model by Proposition 5.3.10. □

Theorem 5.3.12. *If every member of $S \subseteq \mathcal{L}_T$ depends essentially on some set and $\mathbb{S}$ is well-founded on S, then $T \restriction S$ has an ω-model.*

Consider some infinite sequence of sentences $\varphi_1, \varphi_2, \ldots$, where each φ_i is equivalent to the assertion that φ_{i+1} is not true. Although this sequence (i.e. its members) is ungrounded, its set of T-sentences has a standard model. Does the same hold when we are confronted we an infinite sequence in which, say φ_1 refers to $\varphi_2, \varphi_3, \varphi_4$ and φ_2 refers to $\varphi_3, \varphi_4, \varphi_5$ etc. The following proposition, which was jointly proven with Lavinia Picollo, shows that any variation of such a sequence has an acceptable truth assignment.

Theorem 5.3.13. *Let $\Theta = \{\varphi_1, \varphi_2 \ldots\}$ be an infinite set of sentences each of which has a finite dependence set, and assume that the sensitivity-graph of Θ contains no cycle. Then there is an ω-model of $T \restriction \Theta$.*

Proof. Let $\Theta_n = \{\varphi_1, \ldots, \varphi_n\}$ (for $n > 0$). Let Φ_i be the least dependence set of φ_i. For $n > 0$ let

$$\Gamma_n = \{S \subseteq \omega | S \subseteq \bigcup_{i \leqslant n} (\Phi_i \cup \{\varphi_i\}); (\mathbb{N}, S) \vDash T \restriction \Theta_n\}$$

Notice that each Γ_n is finite, because $\bigcup_{i \leqslant n} \Phi_i$ is finite. Moreover, for each $n > 0$ there are $S_1 \subseteq S_2 \subseteq \ldots \subseteq S_n$ with $S_i \in \Gamma_i$ for all $0 < i \leqslant n$. (Proof: For given n, there is some model $(\mathbb{N}, S) \vDash T \restriction \Theta_n$ by Theorem 5.3.11. Choose $S_i := S \cap \bigcup_{k \leqslant i} (\Phi_k \cup \{\varphi_k\})$. Let $j \leqslant i \leqslant n$ be arbitrary. Since $S \cap \Phi_j = S_i \cap \Phi_j$, φ_j has the same truth-value in the model $(\mathbb{N}, S_i)$ as in $(\mathbb{N}, S)$, by Proposition 5.2.2. Moreover, S_i contains φ_j iff S contains it; therefore $T \ulcorner \varphi_j \urcorner$ has the same truth-value in (N, S_i) as in $(\mathbb{N}, S)$. Therefore, $(\mathbb{N}, S_i) \vDash T \ulcorner \varphi_j \urcorner \leftrightarrow \varphi_j$. Thus $S_i \in \Gamma_i$.)

Now we inductively define a finitely branching, infinite tree τ whose nodes are drawn from $\{\varnothing\} \cup \bigcup_{n>0} \Gamma_n$. Let $(\varnothing) \in \tau$. If $(S_0, S_1, \ldots, S_n) \in \tau$ and $S_n \subseteq S_{n+1} \in \Gamma_{n+1}$, then $(S_0, S_1, \ldots, S_n, S_{n+1}) \in \tau$. Nothing else is in τ. Notice that (i) for each n, τ contains a sequence of length n, and (ii) for each sequence $(S_0, S_1, \ldots, S_n) \in \tau$, we have $\varnothing = S_0 \subseteq S_1 \subseteq \ldots \subseteq S_n$ with $S_i \in \Gamma_i$ for $i > 0$. Moreover, because each Γ_i is non-empty and finite, τ is a finitely branching, infinite tree. Thus, by König's lemma, there is an infinite path, providing us with an infinite chain $S_0 \subseteq S_1 \subseteq \ldots$

with $S_i \in \Gamma_i$ for all $i \in \omega$. Let $S = \bigcup_{n\in\omega} S_n$. We will show that $(\mathbb{N}, S) \vDash T \restriction \Theta$. Let φ_i be given. Let $m > i$ be least such that for each S_k with $k \geqslant m$, no sentence in Φ_i will be in S_k unless it was already in S_m. Such an m exists because Φ_i is finite and the S_k are monotone. Clearly, $(\mathbb{N}, S_m) \vDash T\ulcorner\varphi_i\urcorner \leftrightarrow \varphi_i$, because $S_m \in \Gamma_m$ and $i < m$. By choice of m, we have $S \cap (\Phi_i \cup \{\varphi_i\}) = S_m \cap (\Phi_i \cup \{\varphi_i\})$. Since φ_i depends on $\Phi_i \cup \{\varphi_i\}$, $(\mathbb{N}, S)$ is a model of $T\ulcorner\varphi_i\urcorner \leftrightarrow \varphi_i$, by Proposition 5.2.2 and the fact that $(\mathbb{N}, S_m) \vDash T\ulcorner\varphi_i\urcorner \leftrightarrow \varphi_i$. □

Corollary 5.3.14. *If no sentence in S contains a quantifier binding a variable in the scope of the truth predicate and the sensitivity relation restricted to S contains no cycle, then $T \restriction S$ has an ω-model.*

Proof. This follows from the assumption that every member of S has a finite dependence set. □

The above theorems show that sensitivity-graphs provide us with good information about the paradoxicality of those sentences that have essential dependence. However, by Theorem 5.3.4 they must fail to deliver us the relevant information once we are dealing with sentences that lack essential dependence. For example, consider the following version of the Yablo sequence which we may call the *nested* Yablo sequence:

$$Y^*(\overline{n}) \leftrightarrow \exists x > \overline{n} \forall y > x \neg T\ulcorner Y^*(\dot{y})\urcorner$$

The nested Yablo-sequence is just as paradoxical as the original Yablo-sequence: Adding the T-schema for its members to PAT yields a theory that is ω-inconsistent. But for each n, the sensitivity-graph of Y$^*(\overline{n})$ is the empty graph—the very same graph that is also the sensitivity-graph of the most harmless sentences, i.e. of all thoses sentences that do not contain the T-predicate at all. The simple explanation of this phenomenon is that each Y$^*(\overline{n})$ lacks essential dependence: For all $m > n$, Y$^*(\overline{n})$ depends on $\{Y^*(\overline{m}), Y^*(\overline{m+1}), Y^*(\overline{m+2}), \ldots\}$ but it does not depend on the intersection of these sets.

Sensitivity is a good candidate for an aboutness relation as long as we are dealing with sentences that have essential dependence. In particular, the sensitivity concept provides all relevant referential information for those sentences that are usually dealt with in the literature, e.g. the liar, liar cycles, and the Yablo sequence. These sentences contain not more than *one* quantifier binding a variable in the scope of the truth predicate. The distinctive feature of what we dubbed the nested Yablo is that it contains *nested* quantifiers. As we have seen above, the sensitivity-graph fails to contain enough information about paradoxicality as soon as sentences that lack essential dependence are involved. Given this situation, let us see if we can find a generalization of the sensitivity concept that still works in the absence of essential dependence.

5.4. Kripke-games and reference-graphs

In the present section, we will introduce two games for Kripke's fixed-point semantics. We will use the strategies available in these games to derive an infinite system of reference-graphs for each sentence. For sentences that have essential dependence, we can single out a canonical reference-graph—the latter will coincide with the sensitivity-graph of that sentence. The grounding game, which we define in the first subsection, will be used to show that a sentence is grounded if and only if it has a well-founded reference-graph. The verification game, which we introduce in the subsequent section, will be used to give some interesting graph-theoretic descriptions of the Kripke-paradoxical sentences.

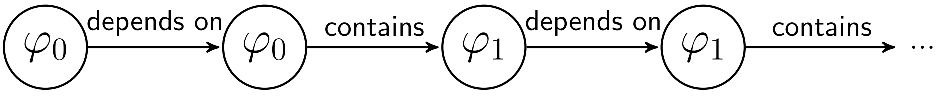
Remark. Games for Kripke's theory of truth have been developed previously, e.g. by Martin [56] and Welch [93]. In the games of Martin and Welch, the players play only with sentences of $\mathcal{L}_T$, while in our games, the players play with sentences and sets of sentences. However, our games enjoy some kind of uniformity: the rules of the games are the same no matter which valuation scheme we are dealing with.

5.4.1. The grounding game

For each sentence φ and set of sentences Φ we define an infinite game of perfect information, the *grounding game* $\mathcal{G}_G(\varphi, \Phi)$ between two players $(\exists)$ and $(\forall)$, such that $(\exists)$ has a winning strategy in $\mathcal{G}_G(\varphi, \Phi)$ iff φ is grounded in Φ. Below we will extract reference-graphs for φ from the strategies of player $(\exists)$ in the game $\mathcal{G}_G(\varphi, \varnothing)$. The rules of $\mathcal{G}_G(\varphi, \Phi)$ are the following:

- $(\forall)$ must move first and choose φ as his first move φ_0.
- As her n-th move $(\exists)$ must choose some set Φ_n on which φ_n depends.
- If $n > 0$, as his n-th move $(\forall)$ must choose some sentence $\varphi_n \in \Phi_n \setminus \Phi$.

The winning conditions for $\mathcal{G}_G(\varphi, \Phi)$ are: $(\exists)$ wins a run of the game if $(\forall)$ cannot move. $(\forall)$ wins a run of the game if it goes on forever.



The most important instances of $\mathcal{G}_G(\varphi, \Phi)$ are those where $\Phi = \varnothing$. We will write $\mathcal{G}_G(\varphi)$ for $\mathcal{G}_G(\varphi, \varnothing)$. For reasons of notational simplicity we will formulate the following definition of a strategy only for the games $\mathcal{G}_G(\varphi)$, but they can be defined for the general case as well.

Call any finite sequence of legal moves in $\mathcal{G}_G(\varphi)$ a *position* of $\mathcal{G}_G(\varphi)$: any position is either an ($\exists$)-position, i.e. a position in which ($\exists$) is to move next, or an ($\forall$)-position. Thus, the set of all $\mathcal{G}_G(\varphi)$ positions forms a tree $\mathrm{T}(\varphi)$ whose nodes (the positions) are ordered by the subsequence relation.

The most important concept of game theory is that of a strategy. We could define a strategy σ for ($\exists$) as a set of rules (or as a function) telling ($\exists$) how to choose her next move in any ($\exists$)-position of the game. But this would be uneconomic in some sense, because we can exclude some position that can never arise as long as ($\exists$) plays according to σ in the first place. Thus, our definition of σ will have a recursive character.

Formally, a *strategy σ for* ($\exists$) *in $\mathcal{G}_G(\varphi)$* is a subtree of $\mathrm{T}(\varphi)$ such that (i) σ is not empty, (ii) if $(\varphi_0, \Phi_0, \ldots, \varphi_n, \Phi_n) \in \sigma$, then for all sentences φ_n such that $\varphi_n \in \Phi_{n-1}$: (φ_0, Φ_0, ..., Φ_{n-1}, φ_n)$\in \sigma$, and (iii) if (φ_0, Φ_0, ..., Φ_{n-1}, φ_n)$\in \sigma$, then for a unique set Φ_n such that φ_n depends on Φ_n: (φ_0, Φ_0,...$\Phi_{n-1}, \varphi_n, \Phi_n$)$\in \sigma$. Thus, σ can be thought of as a partial function defined on the set of ($\exists$)-positions in $\mathcal{G}_G(\varphi)$.

Analogously a *strategy τ for* ($\forall$) can be defined. In this case, roles of (ii) and (iii) are switched and we must allow that there are ($\forall$)-positions p in which τ does not tell ($\forall$) how to move next, namely if p is a winning-position for ($\exists$). A strategy σ is a *winning strategy* for ($\exists$) in $\mathcal{G}_G(\varphi)$ iff she wins every run of $\mathcal{G}_G(\varphi)$ that is *compatible* with σ, i.e. every run of $\mathcal{G}_G(\varphi)$ that is a branch of the tree σ. Informally this means that she wins every run of the game as long as she keeps to the strategy σ, regardless of the moves of her opponent ($\forall$). Analogously a *winning strategy* for ($\forall$) is defined. It is not hard to prove that for any φ the game $\mathcal{G}_G(\varphi)$ is *determined*, that is, either ($\exists$) has a winning strategy in $\mathcal{G}_G(\varphi)$ or ($\forall$) has a winning strategy in $\mathcal{G}_G(\varphi)$.[4] Recall that a set Φ is *D-sound* iff $\Phi \subseteq D(\Phi)$, where D is Leitgeb's dependence operator.

Lemma 5.4.1. *1. A set of sentences Φ is D-sound iff for all $\psi \in \Phi$ there is a set $\Psi \subseteq \Phi$ such that ψ depends on Ψ.*

2. If S is D-sound and $\alpha < \beta$, then $D_\alpha(S) \subseteq D_\beta(S)$.

Proof. 1: Suppose that the right-hand side holds for Φ. Let $\varphi \in \Phi$. By assumption and monotonicity of the dependence relation, φ depends on Φ. Hence $\varphi \in D(\Phi)$. For the reverse direction assume for a contraposition-argument that there is a $\varphi \in \Phi$ such that φ does not depends on any $\Psi \subseteq \Phi$. In particular φ does not depends on Φ. Hence $\varphi \notin D(\Phi)$.

2: The proof is by induction on β. If β=1 then the claim holds trivially by the definition of D-soundness. For β >1 it follows immediately from the induction hypothesis and the first part of the lemma. □

Theorem 5.4.2. *Let S be a D-sound set of sentences. Then φ is grounded in S iff ($\exists$) has a winning strategy in the game $\mathcal{G}_G(\varphi, S)$.*

[4] But notice that for *infinite* games this property is not trivial.

Proof. The proof of the direction from left to right is by induction on the rank_D of a grounded sentence φ, i.e. the least ordinal α such that $\varphi \in D_\alpha(\mathrm{S})$. Let $\text{rank}_D(\varphi) = \alpha$ for some ordinal α. Then φ depends on Φ, for some $\Phi \subset G(S)$ whose members have strictly lower rank_D than φ.

If $\Phi \subseteq S$ then $(\exists)$ can choose S as her first move in $\mathcal{G}_G(\varphi, S)$, and this is a winning strategy for her. Otherwise $\Phi \neq \varnothing$, and by induction hypothesis $(\exists)$ has a winning strategy in $\mathcal{G}_G(\psi, S)$, for all $\psi \in \varphi$. Thus she plays Φ as her first move and whichever $\psi \in \Phi$ $(\forall)$ chooses next, $(\exists)$ simply plays her winning strategy in $\mathcal{G}_G(\psi, S)$. This is a winning strategy for her in $\mathcal{G}_G(\varphi, S)$.

The reverse direction is proved by induction on the strategy-rank of a sentence,

$$\text{rank}_G(\varphi) = \inf\{\text{rank}(\sigma) | \sigma \text{ is a winning-strategy for } (\exists) \text{ in} \mathcal{G}_G(\varphi, S)\}$$

Here, $\text{rank}(\sigma) = \sup\{\ \text{rank}(\tau){+}1)|\ \tau$ is the $(\exists)$-substrategy of σ in $\mathcal{G}_G(\psi, S)$, ψ is a possible response for $(\forall)$ to $(\exists)$'s first move in σ $\}$. Notice that any winning strategy for $(\exists)$ must be well-founded (as a tree), thus $\text{rank}(\sigma)$ is well-defined. Suppose that $(\exists)$ has a winning strategy σ in $\mathcal{G}_G(\varphi, S)$. Let $\text{rank}_G(\varphi) = \alpha$ for some ordinal α. Without loss of generality we may assume that $\text{rank}(\sigma) = \alpha$. Then $\text{rank}_G(\psi) < \alpha$, for all $\psi \in \Psi$, where Ψ is the first move of $(\exists)$ in σ. Thus by induction hypothesis all $\psi \in \Psi$ are grounded in S. Because S is D-sound and φ depends on Ψ, Lemma 5.4.1 (2) yields that φ is grounded in S. (Observe that we could have easily proved $\text{rank}_D(\varphi)$= $\text{rank}_G(\varphi)$ for all grounded sentences φ.) □

The main reason why we are interested in the grounding game is that reference-graphs for φ can be easily extracted from $(\exists)$'s strategies in $\mathcal{G}_G(\varphi)$: Let σ be such a strategy. We define the set of vertices of the *reference-graph of* σ, called $\Gamma(\sigma)$, to be the set of sentences *occurring* in σ. A sentence ψ occurs in σ iff ψ is the last member of a position in the domain of σ, i.e. the last move played by $(\forall)$ leading to this position. Two of its vertices ψ and χ are joined by an arc iff there is an i such that $\psi = \varphi_i$, $\chi = \varphi_{i+1}$ and $(\varphi_0, \Phi_0, \dots \varphi_i, \Phi_i, \varphi_{i+1}) \in \sigma$. Finally, call a graph H *a reference-graph of a sentence* φ iff there is a strategy σ for $(\exists)$ in $\mathcal{G}_G(\varphi)$ such that $H = \Gamma(\sigma)$.

Strategies in the game $\mathcal{G}_G(\varphi)$ tell us something about the semantical properties of the sentence φ. The function $\Gamma(\cdot)$—one can think of it as a kind of projection or forgetful-functor from strategies to graphs—associates to each strategy its graph, that in turn contains some of the strategies semantic information about the sentence.

Theorem 5.4.3. *A strategy σ for $(\exists)$ in $\mathcal{G}_G(\varphi)$ is a winning strategy for $(\exists)$ iff $\Gamma(\sigma)$ is well-founded.*

Proof. A winning strategy for $(\exists)$ is a well-founded tree. □

This, together with Theorem 5.4.2, yields the following result:

Corollary 5.4.4. *A sentence φ is grounded iff φ has a well-founded reference-graph.*

Let us have a closer look at the structure of the set of strategies resp. at the structure of the set of reference-graphs of φ. For ($\exists$)-strategies σ, τ in $\mathcal{G}_G(\varphi)$ let us write $\sigma \preceq \tau$ (σ is a *substrategy of* τ) iff for any position (φ_0, Φ_0, ..., Φ_{n-1}, φ_n) of σ there exists a position (ψ_0, Ψ_0,..., Ψ_{n-1}, ψ_n) of τ such that: if $\varphi_i = \psi_i$ for all i⩽n then $\Phi_i \subseteq \Psi_k$ for all $i \leqslant n$. If for any two such positions one of the $\subseteq$-inclusions is proper, then we write $\sigma \prec \tau$. Accordingly, for graphs G and H let us write $H \preceq G$ iff H is a subgraph of G and $H \prec G$ iff H is a proper subgraph of G. Thus, the set of ($\exists$)-strategies of a sentence φ is partially ordered by $\preceq$, as well as the set of its reference-graphs. Furthermore, $\sigma \preceq \tau$ iff $\Gamma(\sigma) \preceq \Gamma(\tau)$.

Let us call a strategy σ in $\mathcal{G}_G(\varphi)$ (as well as $\Gamma(\sigma)$) *redundant* iff there is a strategy τ in $\mathcal{G}_G(\varphi)$ such that $\tau \prec \sigma$. We call a strategy σ *the canonical strategy* in $\mathcal{G}_G(\varphi)$ (and $\Gamma(\sigma)$ the *canonical reference-graph* of φ) iff σ is not redundant. Clearly, a sentence φ has a canonical reference-graph iff φ has *hereditary essential dependence*, i.e. if φ depends essentially on some set Φ and each member of Φ in turn depends essentially on some set, and so on. For example, the liar sentence has hereditary essential dependence and each member of the (ordinary) Yablo sequence has hereditary essential dependence. As another corollary of Theorem 5.3.4 we obtain

Corollary 5.4.5. *The canonical reference-graph $\Gamma_{min}(\varphi)$ of a sentence φ exist iff the sensitivity-graph of φ is a reference-graph. If the sensitivity-graph of φ is a reference-graph, it is identical to $\Gamma_{min}(\varphi)$.*

Thus, for each φ the set of all reference-graphs of φ is partially orderd by $\preceq$ and has a least element (that might be the empty graph) iff φ has hereditary essential dependence. It has always a largest element, namely the complete graph K_ω, i.e. the graph every sentence occurs as a node of, any two of its nodes being joined by an arc. (The reason is simply that any sentence depends on ω.)

5.4.2. The verification game

The *verification game* $\mathcal{G}_T(\varphi, v, \mathcal{F})$ is quite similar to the grounding game $\mathcal{G}_G(\varphi, \Phi)$, but this time the players are not dealing merely with sentences φ and sets of sentences Φ, but with *facts* (φ, v) and sets of facts $\mathcal{F}$. A *fact*[5] is an ordered pair (φ, v), consisting of a sentence φ and a truth value v that can be either 0 or 1. We let $\mathcal{F}^+ = \{\varphi | (\varphi, 1) \in \mathcal{F}\}$ and $\mathcal{F}^- = \{\varphi | (\varphi, 0) \in \mathcal{F}\}$. Thus, sets of facts are partial interpretations of the truth predicate (in the sense of Kripke) considered as a single set. Therefore, we sometimes identify sets of facts and partial models. We say, for

[5] The notion of fact was first introduced by Yablo [95]. In the sequel paper, we will say more about how Yablo's work relates to ours.

instance, that $\mathcal{F}$ is a sound set of facts, meaning that $\mathcal{F}$ considered as a partial model is sound in the sense of Kripke.[6] A second difference to the grounding game is that a run of the verification game can end in a draw. Before giving a detailed description of the rules of the verification game, let us state a theorem that analogously to Theorem 5.4.2 gives us an idea of what the players ($\exists$) and ($\forall$) are up to in the game (a proof of the theorem is given later):

Theorem 5.4.6. *Let $\mathcal{F}$ be a sound set of facts. Then:*

1. *φ is true in the fixed point of V_L generated by $\mathcal{F}$ iff ($\exists$) has a winning strategy in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$.*
2. *φ is false in the fixed point of V_L generated by $\mathcal{F}$ iff ($\exists$) has a winning strategy in $\mathcal{G}_T(\varphi, 0, \mathcal{F})$.*

To every position of the game $\mathcal{G}_T(\varphi, v, \mathcal{F})$ a *mode* is associated, the mode that a run of the game assumes in this position. This mode is either the *verification mode* or the *falsification mode.* The rules of $\mathcal{G}_T(\varphi, v, \mathcal{F})$ are:

- The game $\mathcal{G}_T(\varphi, 1, \mathcal{F})$ starts in the verification mode, the game $\mathcal{G}_T(\varphi, 0, \mathcal{F})$ starts in the falsification mode.
- ($\forall$) must move first and choose φ as his first move φ_0.
- As her n-th move, ($\exists$) must choose some pair of sets (Φ_n^+, Φ_n^-) such that $\Phi_n^+ \cap \Phi_n^- = \varnothing$, φ_n depends on $\Phi_n^+ \cup \Phi_n^-$, and $Val_{\Phi_n^+}(\varphi_n) = 1$ if the game is in verification mode, and $Val_{\Phi_n^+}(\varphi_n) = 0$ if the game is in falsification mode.
- If $n > 0$, as his n-th move ($\forall$) must choose some sentence $\varphi_{n+1} \in (\Phi_n^+ \setminus \mathcal{F}^+) \cup (\Phi_n^- \setminus \mathcal{F}^-)$. If $\varphi_{n+1} \in \Phi_n^+$ then play continues in the verification mode. If $\varphi_{n+1} \in \Phi_n^-$ then play continues in the falsification mode.

The winning condition for $\mathcal{G}_T(\varphi, v, \mathcal{F})$: If a player cannot move according to the above rules, then the other player wins this run of the game. If a run of the game goes on forever it is declared a draw.

As with the grounding game, we have a special interest in cases where the set parameter $\mathcal{F}$ denotes the empty set; we then write $\mathcal{G}_T(\varphi, v)$. The following definitions are formulated for $\mathcal{G}_T(\varphi, v)$ but apply to the general case as well.

Strategies for the verification game can be defined analogously as for the grounding game. A difference, however, lies in the definition of the positions of the game. An ($\exists$)-position, for example, will not be considered as a sequence of the form (φ_0, Φ_0, ..., Φ_{n-1}, φ_n). Because we want to keep track of the mode of the game, strategies will look like this instead: ((φ_0,v_0), (Φ_0^+, Φ_0^-),...,(φ_{n-1}, v_{n-1}), ($\Phi_{n-1}^+, \Phi_{n-1}^-$),

[6] This notion of soundness must not be confused with our notion of D-soundness.

$(\varphi_n, v_n))$, where v_i is either 1 or 0, according to whether the game is in verification or falsification mode, respectively. Accordingly, we say that a fact (ψ, v) occurs in an $(\exists)$-strategy σ iff (ψ, v) is the last member of a position in the domain of σ. A second difference is that in the verification game there are terminal positions for both player, while in the grounding game it was only $(\forall)$ who could run into a position where moving further was impossible. Keeping these two differences in mind, the notion of a (winning) strategy in the verification game can be defined accordingly, for $(\exists)$ as well as for $(\forall)$. In addition to the notion of a winning strategy, we also have to define what a *non-losing strategy* σ for either player is, namely a strategy that makes sure that she or he wins or draws each run of the game, as long as she or he plays according to σ. The notion of determinacy carries over as well. Either $(\exists)$ has a winning strategy in $\mathcal{G}_T(\varphi, v)$ or $(\forall)$ has a non-losing strategy in $\mathcal{G}_T(\varphi, v)$ and vice versa. The following theorem states an important relationship between the grounding and the verification game.

Theorem 5.4.7. *Let $\mathcal{F}$ be a set of facts. Then player $(\exists)$ has a winning strategy in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$ or in $\mathcal{G}_T(\varphi, 0, \mathcal{F})$ iff $(\exists)$ has a winning strategy in $\mathcal{G}_G(\varphi, \mathcal{F}^+ \cup \mathcal{F}^-)$.*

Proof. $\Rightarrow$: Let $S = \mathcal{F}^+ \cup \mathcal{F}^-$ and assume that $(\exists)$ has a winning strategy σ in $\mathcal{G}_G(\varphi, S)$. By induction on rank$(\sigma) = \sup\{$ rank$(\tau)+1) |\ \tau$ is the $(\exists)$-substrategy of σ in $\mathcal{G}_G(\psi, S)$, ψ is a possible response for $(\forall)$ to $(\exists)$'s first move in $\sigma\}$ we prove that $(\exists)$ has either a winning strategy σ' in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$ or in $\mathcal{G}_T(\varphi, 0, \mathcal{F})$. Notice that rank$(\sigma)$ is well-defined, because as a winning strategy for $(\exists)$ it is a well-founded tree.

Let Ψ be $(\exists)$'s σ-response to φ. Then φ depends on Ψ and by induction hypothesis $(\exists)$ has either a winning strategy σ'_ψ in $\mathcal{G}_T(\psi, 1, \mathcal{F})$ or in $\mathcal{G}_T(\psi, 0, \mathcal{F})$, for all $\psi \in \Psi$. Let Ψ^+ be the set of all members of Ψ such that the first is the case and Ψ^- be the set of all members of Ψ such that the second alternative holds. If $Val_{\Psi^+}(\varphi)$=1, then playing (Ψ^+, Ψ^-) as her first move in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$ followed by σ'_ψ as a repesponse to $(\forall)$'s move ψ is a winning strategy for $(\exists)$ in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$. If $Val_{\Psi^+}(\varphi)$=0, then playing (Ψ^+, Ψ^-) as her first move in $\mathcal{G}_T(\varphi, 0, \mathcal{F})$ followed by σ'_ψ as a response to $(\forall)$'s move ψ is a winning strategy for $(\exists)$ in $\mathcal{G}_T(\varphi, 0, \mathcal{F})$.

Notice that the strategey σ' thus defined is an *orientation* of the strategy σ, i.e. for each $(\exists)$ move Ψ in σ, $\Psi = \Psi^+ \cup \Psi^-$ holds, where (Ψ^+, Ψ^-) is the $(\exists)$-move in σ' that canonically corresponds to Ψ, while the canonically corresponding $(\forall)$ moves in σ respectively σ' are just identical. Moreover, σ' is the only orientation of σ that is a winning-strategy for $(\exists)$ in either $\mathcal{G}_T(\psi, 1, \mathcal{F})$ or in $\mathcal{G}_T(\psi, 0, \mathcal{F})$.

$\Leftarrow$: Suppose has $(\forall)$ a winning strategy τ in $\mathcal{G}_G(\varphi, \mathcal{F}^+ \cup \mathcal{F}^-)$. Consider the following $(\forall)$-strategy τ': If (Φ^+, Φ^-) is a move by $(\exists)$ in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$, then play the τ-response to $\Phi^+ \cup \Phi^-$. This is a winning strategy in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$ for $(\forall)$, because τ' must be an endless tree: τ' is an orientation of τ and thus canonically isomorphic to τ, and τ as a $(\forall)$-winning in the grounding game is an endless tree. □

Let us define a dependence relation between facts: (φ, v) depends on $\mathcal{F}$ iff $(\varphi,\ v) \in \mathrm{V}_L(\mathcal{F})$. The following Lemma as well as its proof are completely analogous to Lemma 5.4.1.

Lemma 5.4.8. *A consistent set of facts $\mathcal{F}$ is sound iff for all $f \in \mathcal{F}$ there is a set $\mathcal{E} \subseteq \mathcal{F}$ such that f depends on $\mathcal{E}$.*

In short, a set of facts is sound iff it is consistent and closed under the dependence relation between facts.[7]

Now we can turn to the

Proof of Theorem 5.4.6. $\Rightarrow$: Suppose φ is true in the fixed point of V_L generated by $\mathcal{F}$. Hence φ is grounded in $\mathcal{F}^+ \cup \mathcal{F}^-$ and by Theorem 5.4.2, $(\exists)$ has a winning strategy σ in $G(\varphi, \mathcal{F}^+ \cup \mathcal{F}^-)$. Then the strategy σ' as defined in the proof of Theorem 5.4.7 is a winning strategy for $(\exists)$ in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$.

$\Leftarrow$: Suppose $(\exists)$ has a winning strategy σ' in $\mathcal{G}_T(\varphi, 1, \mathcal{F})$. Then by Theorem 5.4.7 $(\exists)$ has a winning strategy σ in $\mathrm{G}(\varphi,\ \mathcal{F}^+ \cup \mathcal{F}^-)$. Thus by Theorem 5.4.2, φ is grounded in $\mathcal{F}^+ \cup \mathcal{F}^-$. Using Lemma 5.4.8 one shows by induction on the D-rank of φ that φ is true in the fixed-point of V_L generated by $\mathcal{F}$.

The proof of the second part of 5.4.6 is completely analogous.

Now let us see how we can extract *signed reference-graphs* from verification-strategies. The first step is to define the *unsigned reference-graph* $\Gamma(\sigma)$ of a $(\exists)$-strategy in $\mathcal{G}_T(\varphi, v)$, completely analogous to the grounding-game: A sentence φ is a vertex of $\Gamma(\sigma)$ iff for $v \in \{0, 1\}$ the fact (φ, v) occurs in σ. Two of its vertices ψ and χ are joined by an arc in the reference-graph iff for $v_\psi,\ v_\chi \in \{0, 1\}$ the fact (ψ, v_ψ) occurs in some position of σ that is assigned to a set of facts which contains (χ, v_χ) as an element. So much for $\Gamma(\sigma)$.

The *signed reference-graph* $\Gamma^+_-(\sigma)$ of σ is obtained from $\Gamma(\sigma)$ by labeling its arcs according to the following rules: First, if there are $v_\psi = v_\chi \in \{0, 1\}$ such that (ψ, v_ψ) occurs in some position of σ that is assigned to a set of facts which contains (χ, v_χ) as an element, the we put the label '+' on the arc determined by the ordered pair $((\psi, v_\psi),\ (\chi, v_\chi))$. Second, if there are $v_\psi \neq v_\chi \in \{0, 1\}$ such that (ψ, v_ψ) occurs in some position of σ that is assigned to a set of facts which contains (χ, v_χ) as an element, then we put the label '-' on the arc determined by the ordered pair $((\psi, v_\psi),$ $(\chi, v_\chi))$.

In this way at least one label is assigned to every arc of $\Gamma(\sigma)$, but unfortunately we cannot exclude that there are cases where both '+' and '-' are put on the same arc. Let us call the strategy σ *incoherent* if this is the case and *coherent* otherwise.

[7] Notice that for set of facts being 'closed under dependence' is some kind of 'downward-closedness' while being closed under the operator V_L is some kind of 'upward-closedness'. Observe that for any $(\exists)$-strategy σ in the verification game the set of facts $(\Phi^+_\sigma,\ \Phi^-_\sigma)$ occurring in σ is closed under the dependence relation by the definition of a strategy.

Incoherent strategies, awkward as they might be, pose, however, not too much of a threat to our definition of a signed reference-graphs. Analogously as we have done for strategies in the grounding-game, we can define a substrategy-relation $\preceq$ on the set of (∃)-strategies in $\mathcal{G}_T(\varphi, v)$, comparing sets of facts instead of sets of sentences occurring in the positions of the strategies. We can prove that for each strategy τ there is always a strategy $\sigma \preceq \tau$ such that σ is coherent. This means that incoherence is a phenomenon due to redundency and that we can always cut out the incoherent part, obtaining a coherent substrategy. Because a substrategy $\sigma \preceq \tau$ contains more semantic information about the sentence φ than τ anyway, we can completely forget about incoherent strategies and redefine *strategy* as *coherent strategy*.

The introduction of signed reference-graphs seems to us to be of great important for a full understanding of the notion of paradoxicality. Obviously, the canonical (unsigned) reference-graph of the liar is identical with the canonical (unsigned) reference-graph of the truth-teller; but while the liar is clearly paradoxical, the truth-teller is not. As Herzberger pointed out, both sentences suffer from some form of semantic regress—they are both pathological—, and this is captured in our framework by the fact that all of their (unsigned) reference-graphs contain a directed cycle as a subgraph. However, what distinguishes the liar from the truth-teller, and causes the former to be paradoxical, is that the liar involves some kind of negative self-reference, while the truth-teller only exhibits some kind of positive self-reference.

We remark *en passant* that more parameters should be taken into account for a full understanding of the paradoxes. For example, we believe that another thing of utmost importance is *where* a cycle occurs in a reference-graph. For example, suppose that the reference-graph of φ contains a directed cycle C as a subgraph. One might speculate that φ is dangerous only if φ is a vertex of the subgraph C. However, we have to leave these considerations open for future research.

5.4.3. Kripke-paradoxicality

Strategies in the verification game are more complex than strategies in the grounding game, containing more information about the semantical properties of a sentence. Aside from the nice characterization of the groundedly true sentences by Theorem 5.4.6, the main reason why we are interested in the verification game is that we can also give a characterization of the set of the Kripke-paradoxical sentences just in terms of strategies of the verification game. Call a (∃)-strategy σ in $\mathcal{G}_T(\varphi, v)$ *consistent* iff the set of facts $\mathcal{F}_\sigma$ occurring in σ is consistent, i.e iff $\mathcal{F}_\sigma^+ \cap \mathcal{F}_\sigma^- = \varnothing$. Let us call a strategy for (∃) *faithful* iff it is a non-losing strategy.

Theorem 5.4.9. *A sentence φ is has a definite truth value v in some fixed-point (Φ^+, Φ^-) of V_L iff there is a consistent faithful strategy σ for player (∃) in $\mathcal{G}_T(\varphi, v)$.*

Proof. $\Rightarrow$: Suppose φ has the truth value v in some fixed point $\mathcal{F}$ of V_L. By Lemma 5.4.8 for each $f \in \mathcal{F}$ there is a set $\mathcal{E} \subseteq \mathcal{F}$ such that ψ depends on $\mathcal{E}$. Thus, using the Axiom of Choice, we can build up a strategy σ for player $(\exists)$ in $\mathcal{G}_T(\varphi, v)$ layer by layer. σ is faithful by construction and consistent because $\mathcal{F}$ is consistent by definition.

$\Leftarrow$: Let σ be a consistent faithful strategy for player $(\exists)$ in $\mathcal{G}_T(\varphi, v)$. Let $\mathcal{F}_\sigma$ be the set of all facts occurring in σ. Then $\mathcal{F}_\sigma$ is consistent. Because σ is a faithful strategy, for each $f \in \mathcal{F}_\sigma$ there is a set $\mathcal{E} \subseteq \mathcal{F}_\sigma$ such that ψ depends on $\mathcal{E}$. Hence by Lemma 5.4.8, $\mathcal{F}_\sigma$ is sound. Hence there is some fixed point (Φ^+, Φ^-) of V_L extending $\mathcal{F}_\sigma$ and φ has the truth-value v in (Φ^+, Φ^-). □

Hence, a sentence φ is Kripke-paradoxical (with respect to V_L) iff any strategy for $(\exists)$ in $\mathcal{G}_T(\varphi, 1)$ or in $\mathcal{G}_T(\varphi, 0)$ is either a losing-strategy for $(\exists)$—i.e. a strategy that can defeated by some $(\forall)$-strategy—or inconsistent.

Now let us apply our machinery to gain some information on the Kripke-paradoxical sentences. We need the following preliminary theorem.

Theorem 5.4.10. *Let σ be a strategy for $(\exists)$ in $\mathcal{G}_T(\varphi, v)$. Then there is a definite truth value v^* and a faithful strategy σ^* for $(\exists)$ in $\mathcal{G}_T(\varphi, v^*)$ such that $\Gamma(\sigma^*) = \Gamma(\sigma)$.*

For a proof of this deep theorem I refer the reader again to Beringer & Schindler [7].

As a consequence, there are large classes of sentences that can easily be recognized as not being Kripke-paradoxical (with respect to V_L):

We call a signed reference-graph *positive* iff it has positive arcs only.[8]

Theorem 5.4.11. *If a sentence φ has a faithful*[9] *positive signed reference-graph, then φ is not Kripke-paradoxical.*

Theorem 5.4.12. *If a sentence φ has a reference-graph which is a tree, then φ is not Kripke-paradoxical.*

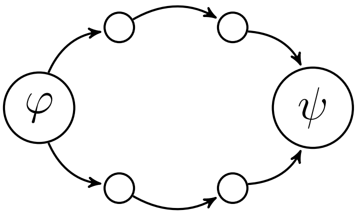
Proof. Let σ be a strategy in σ for player $(\exists)$ in $\mathcal{G}_T(\varphi, v)$, for some definite truth value v, such that $\Gamma(\sigma)$ is a tree. By Lemma 5.4.10 we can assume that σ is faithful. Because $\Gamma(\sigma)$ is a tree no sentence ψ occurring in σ can occur in both contexts $(\psi, 1)$ and $(\psi, 0)$. Thus σ is consistent. By Lemma 5.4.9 φ has the definite truth value v in some fixed point. □

If a reference-graph is not a tree then it either contains a directed cycle as a subgraph or it contains a type of graph as a subgraph that we may call a *double path*, i.e. a

[8] Accordingly, call a signed graph *negative* iff every arc bears the symbol '-'. Examples: The unique orientation of the liar-graph is negative. The unique orientation of the truth-teller-graph is positive. In the unique orientation of the ordinary Yablo-graph every arc is negative.

[9] A signed reference-graph is *faithful* if its strategy is faithful.

graph consisting of two paths originating both from the same vertex and rejoining in a different vertex, not touching each other in between.[10]



A double path between φ and ψ

Thus, we obtain:

Theorem 5.4.13. *If a sentence φ is Kripke-paradoxical, then for each reference-graph Γ of φ at least one of the following holds:*

1. *Γ contains a directed cycle.*
2. *Γ contains a double path.*

It is worth noticing that while the directed cycle is the reference pattern underlying the liar family, the double path is the reference pattern of any member of the Yablo sequence. However, it can be shown that if φ has a reference graph with no cyles and only *finitely* many double arcs, then φ is not Kripke-paradoxical. Unlike cycles, double arcs must come in flocks in order to make a reference-graph dangerous. In our paper [7] we state a conjecture according to which every reference-graph of a Kripke-paradoxical sentence is reducible either to the liar or Yablo-graph.

[10] The reason for this is that reference-graphs emerge from directed trees (namely strategies) by some collapsing operation.

6. Axioms for grounded truth

A classical untyped theory of truth cannot contain all instances of the T-schema nor can it contain all compositional principles, on pain of contradiction. Quite a few authors (e.g. Horwich [47], Restall [77]) have suggested that we ought to accept those principles at least for all *grounded* sentences. The notion of groundedness can be explicated by any of Kripke's minimal fixed point theories. Kripke's theories of truth are *semantic*. They provide definitions of truth in a metalanguage (usually set theory, but certain subsystems of second-order arithmetic suffice.) This is not quite what we need. The main purpose of the truth predicate is to enable us to express and reason with generalizations. Although we can derive semantic consequence relations from Kripke's models, such consequence relations are too complex to be useful in actual reasoning, because of the complexity of the models. This prompts the search for axiomatizations of Kripke's theories of truth. Given the complexity of Kripke's construction, such axiomatizations cannot be complete. Thus, what we are aiming at are axiomatic theories that are sound with respect to the models in question and that capture important features of them.

6.1. KF and VF

Feferman [20] has given an axiomatization of the Strong Kleene fixed points, while Cantini [11] has provided an axiomatization of the supervaluational fixed points.

Definition 6.1.1. The system KF (the acronym stands for 'Kripke-Feferman') is obtained from PAT by adding the following 13 axioms:

1. $\forall s \forall t (T(s \underset{\cdot}{=} t) \leftrightarrow s^\circ = t^\circ)$
2. $\forall s \forall t (T(s \underset{\cdot}{\neq} t) \leftrightarrow s^\circ \neq t^\circ)$
3. $\forall t (T \underset{\cdot}{T} t \leftrightarrow T t^\circ)$
4. $\forall t (T \underset{\cdot}{\neg} \underset{\cdot}{T} t \leftrightarrow (T \underset{\cdot}{\neg} t^\circ \vee \neg Sent_T(t^\circ)))$
5. $\forall x (Sent_T(x) \rightarrow (T \underset{\cdot}{\neg} \underset{\cdot}{\neg} x \leftrightarrow T x))$
6. $\forall x \forall y (Sent_T(x \underset{\cdot}{\wedge} y) \rightarrow (T(x \underset{\cdot}{\wedge} y) \leftrightarrow T x \wedge T y))$

7. $\forall x \forall y(Sent_T(x \underset{\cdot}{\wedge} y) \to (T \underset{\cdot}{\neg}(x \underset{\cdot}{\wedge} y) \leftrightarrow T \underset{\cdot}{\neg} x \vee T \underset{\cdot}{\neg} y))$

8. $\forall x \forall y(Sent_T(x \underset{\cdot}{\vee} y) \to T(x \underset{\cdot}{\vee} y) \leftrightarrow Tx \vee Ty))$

9. $\forall x \forall y(Sent_T(x \underset{\cdot}{\vee} y) \to (T \underset{\cdot}{\neg}(x \underset{\cdot}{\vee} y) \leftrightarrow T \underset{\cdot}{\neg} x \wedge T \underset{\cdot}{\neg} y))$

10. $\forall x \forall v(Sent_T(\underset{\cdot}{\forall} vx) \to (T(\underset{\cdot}{\forall} vx) \leftrightarrow \forall t T(x(t/v)))$

11. $\forall x \forall v(Sent_T(\underset{\cdot}{\forall} vx) \to (T(\underset{\cdot}{\neg} \underset{\cdot}{\forall} vx) \leftrightarrow \exists t T(\underset{\cdot}{\neg} x(t/v))))$

12. $\forall x \forall v(Sent_T(\underset{\cdot}{\exists} vx) \to (T(\underset{\cdot}{\exists} vx) \leftrightarrow \exists t T(x(t/v))))$

13. $\forall x \forall v(Sent_T(\underset{\cdot}{\exists} vx) \to (T(\underset{\cdot}{\neg} \underset{\cdot}{\exists} vx) \leftrightarrow \forall t T(\underset{\cdot}{\neg} x(t/v))))$

The system KF can be seen as being obtained by turning the inductive clauses in the definition of the Strong Kleene fixed points into axioms (for more on that cf. Halbach [38, p. 202-204]).

(Cons) is the following principle:

$$\forall x(Sent_T(x) \to (T \underset{\cdot}{\neg} x \to \neg Tx))$$

Proposition 6.1.2 (Cantini, Halbach)**.** $\mathsf{KF} + (Cons) = \mathsf{KF} + T - Out.$

KF+(Cons) does not prove more arithmetical theorems than KF alone (cf. Cantini [10]). KF has a very nice adequacy property:

Proposition 6.1.3. *Let S be a partial model. Then $S = \mathcal{J}_{SK}(S)$ iff $(\mathbb{N}, S^+) \vDash \mathsf{KF} + (\forall x(Tx \to Sent_T(x)))$.*

(Recall that S^+ is the extension of the truth predicate in the partial model S; so KF somehow captures the closed-off models.) For a proof of the above proposition, see e.g. Halbach [38]. Feferman has calibrated the proof-theoretic strength of KF:

Theorem 6.1.4 (Feferman [20])**.** KF *proves the same arithmetical sentences as* RA_{ϵ_0}.

This result is probably a bit disappointing: Although KF+(Cons) proves T-Out and thereby satisfies the generalizing function of truth (see section 4.1), its deductive power is less than that of the Tarskian hierarchy. KF is able to define the truth predicates of RT up to the level ϵ_0 (as Halbach [38, chap. 15.3] has shown) but it cannot recover the whole Tarskian hierarchy. However, in a sense it is not the truth-theoretic axioms of KF that are to blame but rather the fact that PAT does not contain enough transfinite induction. If one adds more induction to KF, the extended system can define more levels of the Tarskian hierarchy. Cantini's system VF, on the other hand, is impredicative and therefore much stronger than RT_{ϵ_0}.

Definition 6.1.5. The system VF (the acronym stands for 'Van Frassen') is obtained from PAT by adding the following axioms:

1. $\forall x_1 \dots \forall x_n(T\ulcorner\varphi(\dot{x}_1, \dots, \dot{x}_n)\urcorner \to \varphi(x_1, \dots, x_n))$
2. $\forall s\forall t(T(s\underset{\cdot}{=}t) \leftrightarrow s^\circ = t^\circ)$
3. $\forall s\forall t(T(s\underset{\cdot}{\neq}t) \leftrightarrow s^\circ \neq t^\circ)$
4. $\forall x(Sent_T(x) \wedge Prov_{PAT}(x) \to Tx)$
5. $\forall x\forall v(Sent_T(\underset{\cdot}{\forall} vx) \to (\forall tT(x(t/v)) \to T(\underset{\cdot}{\forall} vx)))$
6. $\forall x(Tx \to T\ulcorner T\dot{x}\urcorner)$
7. $\forall x(T\underset{\cdot}{\neg}\underset{\cdot}{T}\dot{x} \to T\underset{\cdot}{\neg}x)$
8. $\forall x\forall y(Sent_T(x\underset{\cdot}{\wedge}y) \to (T(x\underset{\cdot}{\to}y) \to (Tx \to Ty)))$
9. $\forall x(T(\ulcorner T\dot{x} \to \neg T\underset{\cdot}{\neg}\dot{x}\urcorner))$

In contrast to KF, the axioms of VF do not mirror the inductive process which generates the fixed points. Several authors have criticised VF because its axioms seem somewhat unrelated or arbitrary. Halbach has remarked (private communication) that the axioms of VF rather look like axioms of a modal theory than those of a truth theory (for example, axiom 5 is just the predicate analogon of the Barcan formula, axiom 6 that of the S4 axiom, axiom 8 that of the K axiom, axiom 4 is a restricted version of the necessitation rule etc.). Nevertheless, VF captures the supervaluational fixed points to a certain extent:

Theorem 6.1.6 (Cantini [11])**.** *Let S be a partial model. If $S = \mathcal{J}_{FV}(S)$ is consistent, then* $(\mathbb{N}, S^+) \vDash \mathsf{VF}$.

The converse does not hold, due to complexity of the underlying jump operator.

Theorem 6.1.7 (Cantini [11])**.** VF *proves the same arithmetical sentences as* ID_1.

The system ID_1 has the same proof-theoretic ordinal (namely, the so-called Bachmann-Howard ordinal[1]) as the system $\Pi^1_1\text{-}\mathsf{CA}_0^-$ (where the minus indicates that no set parameters are allowed in the comprehension axioms). They are both impredicative theories.[2]

In what sense can KF or VF be viewed as theories of *grounded* truth? As we have seen in the last chapters, a set of grounded sentences is generated by an inductive

[1] Cf. Pohlers [66, chap. 9].

[2] For a definition of the system ID_1, see the appendix.

process: one starts with a certain set of sentences—say, the set of T-free sentences or the set of T-free sentences plus all tautologies in the full language $\mathcal{L}_T$—and then one repeatedly adds sentences that semantically depend on (or are grounded in) the sentences of the previous level until a fixed point is reached. A set generated in this way is closed under certain operations (depending on the chosen valuation scheme); normally, they will be closed under (introduction of) negation, conjunction, disjunction etc. At first sight, neither KF nor VF seem to mirror these closure conditions. However, Feferman [20, p. 20] noticed the following. One can define a predicate, $G(x)$, by stipulating

$$G(x) \leftrightarrow (Sent_T(x) \wedge (Tx \vee T\underset{.}{\neg}x) \wedge \neg(Tx \wedge T\underset{.}{\neg}x))$$

Then the following principles are provable from KF:

1. $G(x) \rightarrow G(\underset{.}{\neg}x)$
2. $G(x) \rightarrow G(\underset{.}{T}x)$
3. $G(x) \wedge G(y) \rightarrow G(x\underset{.}{\wedge}y)$
4. $G(x) \wedge G(y) \rightarrow G(x\underset{.}{\vee}y)$
5. $\forall t G(x(t/v)) \rightarrow G(\underset{.}{\forall}vx))$

Thus, one can *derive* from KF axioms that mirror the closure conditions of the set of grounded sentences. Feferman also mentions that, instead of using a system like KF, one could work in a language containing besides a truth predicate T also a primitive grounding predicate G, which is then axiomatized by principles such as above. In [20], he is reluctant to do so, saying that the resulting system would be "formally weaker", but in [21] he actually proposed such a theory, called DT (the acronym stands for 'determinate truth'). DT contains besides grounding axioms all the compositional truth axioms relativized to G. For example, instead of

$$Tx\underset{.}{\wedge}y \leftrightarrow Tx \wedge Ty$$

one has

$$G(x) \wedge G(y) \rightarrow (Tx\underset{.}{\wedge}y \leftrightarrow Tx \wedge Ty)$$

Simultaneous axiomatizations of groundedness and truth seem to be an interesting way of axiomatizing Kripke fixed points.

First, one might find a more 'uniform' method of axiomatizing minimal Kripke fixed points; one just has to look at the closure conditions that a certain fixed point satisfies (usually, these closure conditions can be easily read off of the truth tables of the valuation scheme in question) and then turn these conditions into axioms for groundedness. In particular, in this way we might be able to find an axiomatization for Leitgeb's theory of truth.

Second, this approach might be interesting for disquotationalists. KF and VF are compositional theories of truth and have therefore been rejected by disquotationalists such as Horwich. Nevertheless, Horwich flirts with the idea of restricting the T-schema to grounded sentences in order to block the liar paradox.

> The intuitive idea is that an instance of the equivalence schema will be acceptable, even if it governs a proposition concerning truth [...], as long as that proposition is *grounded* [...] ([47, p. 81])

And

> A well-known worked-out approach based on the notion of grounding is given in Saul Kripke's "Outline of a Theory of Truth" [...], but in a way that invokes Tarski-style compositional principles. The present suggestion is that such principles can be avoided, offering a solution that squares with minimalism. ([47, p. 82, fn 11])

Now, one possiblity would be to give axioms for the predicate G and then to adopt, in addition, the relativized T-schema

$$G(\ulcorner\varphi\urcorner) \rightarrow (T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi)$$

Similar ideas can also be found in the writings of other authors. For example, Leitgeb begins his paper by raising the following question: "What kinds of sentences with truth predicate may be inserted plausibly and consistently into the T-scheme? We state an answer in terms of dependence: those sentences which depend directly or indirectly on non-semantic states of affairs (only)." ([55, p. 155]) See also the discussion by Field [26, chapt. 7.4 and 7.6].

6.2. Simultaneous axiomatizations of groundedness and truth

In this section we suggest a list of possible axioms for a theory of grounded truth. We start with an axiom that identifies the concept of grounding in terms of truth:

G0 $\forall x(Sent_T(x) \rightarrow (G(x) \leftrightarrow (Tx \vee T\underset{\cdot}{\neg}x)))$

This axiom says that the grounded sentences are exactly those that are true or have a true negation. This is exactly the definition of grounding that Kripke gave: truth comes first, and grounding is derived. Thus the predicate G is eliminable. However, on Leitgeb's approach the order is reversed: grounding is conceptually prior to truth.

First the grounding hierarchy is defined, then the truth hierarchy is defined based on the grounding hierarchy. Whether one adopts the predicate G as a primitive or considers it as an abbreviation according to G0 won't play a huge role in our investigations, but let us assume for definiteness that G is a primitive.

We continue with some base and closure axioms. These mirror (1)-(7) of Proposition 5.2.6 stated earlier for Leitgeb's theory. We discuss the axioms in more detail after we have presented all of them.

Base Axioms.

G1 $\forall x(Sent_{PA}(x) \rightarrow G(x))$
G2 $\forall x(Sent_T(x) \rightarrow (Prov_{PAT}(x) \rightarrow G(x)))$

Closure Axioms.

G3 $\forall x(Sent_T(x) \rightarrow (G(\underset{\cdot}{T}x) \leftrightarrow G(x)))$
G4 $\forall x(Sent_T(x) \rightarrow (G(x) \leftrightarrow G(\underset{\cdot}{\neg}x)))$
G5 $\forall x\forall y(Sent_T(x\underset{\cdot}{\vee}y) \rightarrow (G(x) \wedge G(y) \rightarrow G(x\underset{\cdot}{\vee}y)))$
G6 $\forall x\forall y(Sent_T(x\underset{\cdot}{\wedge}y) \rightarrow (G(x) \wedge G(y) \rightarrow G(x\underset{\cdot}{\wedge}y)))$
G7 $\forall x\forall v(Sent_T(\underset{\cdot}{\forall}vx) \rightarrow (\forall tG(x(t/v)) \rightarrow G(\underset{\cdot}{\forall}vx)))$

Jump Axiom.

Let $Rel(x, y)$ represent the relation that holds between a closed $\mathcal{L}_T$-formula φ and a (finite) sequence of $\mathcal{L}_{PA}$-formulae $\langle\psi_1(x), \dots, \psi_n(x))\rangle$ iff every subformula of φ of the form Tt occurs in the context $\psi_i(t) \wedge Tt$ within φ for some $i \leqslant n$. We write $\forall\sigma$ instead of $\forall x(seq(x) \rightarrow \dots)$, where $seq(x)$ expresses that x is (the code of) a sequence. We write $lh(\sigma)$ for the length of σ, and $\sigma(i)$ for the i-th member of the sequence σ.

G8 $\forall x\forall\sigma(Rel(x, \sigma) \wedge \forall i < lh(\sigma)\forall v(Sat(v, \sigma(i)) \rightarrow G(v)) \rightarrow G(x))$

Here, $Sat(x, y) := T\underset{\cdot}{s}(y, x)$, where $\underset{\cdot}{s}$ represents the function s that applied to the code of a formula $\varphi(x)$ and a number n yields the code of $\varphi(\overline{n})$ (cf. section 2.1).[3] In a more readable form, G8 simply says that

$$\text{If for all } i \leqslant n,\ \psi_i \subseteq G, \text{ then } \varphi^{(\psi_1,\dots,\psi_n)} \text{ is grounded}$$

[3] It would be more natural just to relativize (in the usual sense) all quantifiers to φ; however, the present formulation is more convenient for the proof of proposition 6.3.5 below. Given an axiom to the effect that G is closed under arithmetical equivalence (as considered in remark 6 below) both formulations would amount to the same.

where $\varphi^{(\psi_1,\dots,\psi_n)}$ is the result of relativizing φ to the sequence $(\psi_1,\dots,\psi_n)$.

Axioms for Conditional Dependence.

G9 $\forall x \forall y(Sent_T(x \underset{\cdot}{\vee} y) \to (T(x) \to G(x \underset{\cdot}{\vee} y)))$
G10 $\forall x \forall y(Sent_T(x \underset{\cdot}{\wedge} y) \to (T(\underset{\cdot}{\neg} x) \to G(x \underset{\cdot}{\wedge} y)))$
G11 $\forall x \forall y(Sent_T(x \underset{\cdot}{\vee} y) \to (T(x \underset{\cdot}{\vee} y) \wedge T(\underset{\cdot}{\neg} x) \to G(y)))$
G12 $\forall x \forall y(Sent_T(x \underset{\cdot}{\wedge} y) \to (T(x \underset{\cdot}{\wedge} y) \to G(x) \wedge G(y)))$

Axioms of Truth

T1 $\forall s \forall t(T(s \underset{\cdot}{=} t) \leftrightarrow s^{\circ} = t^{\circ})$
T2 $\forall t(G(t^{\circ}) \to (T\underset{\cdot}{T}t \leftrightarrow Tt^{\circ}))$
T3 $\forall x(Sent_T(x) \to (G(x) \to (T\underset{\cdot}{\neg} x \leftrightarrow \neg Tx))$
T4 $\forall x \forall y(Sent_T(x \underset{\cdot}{\wedge} y) \to (G(x) \wedge G(y) \to (T(x \underset{\cdot}{\wedge} y) \leftrightarrow Tx \wedge Ty))$
T5 $\forall x \forall y(Sent_T(x \underset{\cdot}{\vee} y) \to (G(x) \wedge G(y) \to (T(x \underset{\cdot}{\vee} y) \leftrightarrow Tx \vee Ty))$
T6 $\forall x \forall v(Sent_T(\underset{\cdot}{\forall} vx) \to (G(\underset{\cdot}{\forall} vx) \to (T(\underset{\cdot}{\forall} vx) \leftrightarrow \forall t T(x(t/v))))$
T7 $\forall t_1 \dots \forall t_n(G\ulcorner \varphi(\underset{\cdot}{t_1},\dots,\underset{\cdot}{t_n})\urcorner \to (T\ulcorner \varphi(\underset{\cdot}{t_1},\dots,\underset{\cdot}{t_n})\urcorner \leftrightarrow \varphi(t_1^{\circ},\dots,t_n^{\circ})))$

Axiom G1 states that all T-free sentences are grounded; this axiom is satisfied in any Kripke fixed point. Axiom G2 states that all theorems of PAT are grounded. This captures the thought that everything that is already decided by the base theory is grounded. We could have added that also all sentences that are refuted by the base theory are grounded, but this is redundant, because it will follow from the negation axiom. G2 is sound with respect to the supervaluations and the Leitgeb valuation scheme, but it is not sound with respect to the Kleene schemes, since, for example, the disjunction $\lambda \vee \neg\lambda$ is neither grounded under the Strong nor the Weak Kleene scheme (where λ is again the Liar sentence).

Axioms G3−G7 give closure conditions. This is why most of them have the form of a conditional rather than a biconditional. The result of adding the right-to-left direction of G5−G7 is not sound with respect to the Leitgeb and supervaluations scheme. For example, $\lambda \vee \neg\lambda$ (where λ is the Liar sentence) is grounded relative to the Leitgeb and the supervaluational scheme, but none of its disjuncts is. In the case of Strong Kleene, one might add axioms such as:

$$\forall x \forall y(Sent_T(x \underset{\cdot}{\vee} y) \to (G(x \underset{\cdot}{\vee} y) \to G(x) \vee G(y)))$$

If one intends to axiomatize Kripke's construction with the Weak Kleene scheme, then one might add even stronger axioms like:

$$\forall x \forall y(Sent_T(x \underset{\cdot}{\vee} y) \to (G(x \underset{\cdot}{\vee} y) \leftrightarrow G(x) \wedge G(y))) \qquad (\mathrm{G5}^{c})$$

Notice that G5^c + G2 + T7 is inconsistent over PAT.

The idea underlying the jump axiom G8 is that a sentence that attributes truth or falsity to some subset(s) of G is itself grounded. For example, if we already know that the predicate $\psi(x)$ holds only of grounded sentences, then intuitively statements like $\exists x(\psi(x) \wedge Tx)$ are grounded too. In the statement of the axiom, the formulae ψ_i must be arithmetical (i.e. T-free), since otherwise an inconsistency would occur. Notice that G8 is not sound with respect to the Weak Kleene scheme (roughly, the conditions for a quantified statement to be true are so strong under WK that they don't allow for relativization).

As pointed out (cf. section 5.2), axioms G9−G12 are not satisfied by the Leitgeb's scheme; rather, they apply only to the Strong Kleene and supervaluations scheme.

Notice that replacing $G(x) \wedge G(y)$ by $G(x \dot{\vee} y)$ in the antecedent of T5 would be unsound with respect to the Leitgeb and supervaluations scheme. For example, while $\lambda \vee \neg\lambda$ is grounded relative to the Leitgeb and the supervaluations scheme, none of its disjuncts is. For similar reasons, the antecedent of T4 can not be replaced by $G(x \dot{\wedge} y)$. For this reason, T7 cannot be derived from T1−T6. However, given the Strong or Weak Kleene schema, such a replacement is indeed sensible. In that event, T7 will be derivable and does not have to be assumed as an axiom.

One might consider the idea of adding a further axiom that states that G is closed under PAT-equivalence:

$$\forall x \forall y (Prov_{PAT}(x \dot{\leftrightarrow} y) \wedge G(x) \rightarrow G(y))$$

This is indeed sound with respect to the Leitgeb and the supervaluational scheme, but it does not hold under any of the Kleene schemes. The above formula would give a more unified picture, but would not add anything to the proof-theoretic strength.

Let us quickly present a fairly simple model of G0 + G1 + G3-G8 + T1-T7. Notice that models now have the form $(\mathbb{N}, X, Y)$, where $X \subseteq \omega$ interprets G and $Y \subseteq \omega$ interprets T. Again, we identify sentences with their codes.

Let $\Sigma_0 := \mathcal{L}_{PA}$, and let $\Sigma_{\alpha+1} \subseteq Fm_T$ be the smallest superset of Σ_α such that (i) whenever $\varphi, \psi \in \Sigma_\alpha$, then $\varphi \wedge \psi, \varphi \vee \psi, \neg\varphi, T\ulcorner\varphi\urcorner \in \Sigma_{\alpha+1}$, (ii) whenever $\varphi(t) \in \Sigma_\alpha$ for all t, then $\forall x \varphi \in \Sigma_{\alpha+1}$, (iii) whenever $\psi_1(x), \ldots, \psi_n(x)$ are T-free formulae each of which (elementarily) defines a subset of Σ_α and φ is relativized (in the above sense) to the sequence of $\psi_i(x)$'s, then $\varphi \in \Sigma_{\alpha+1}$. At limit points, we take unions. Let Σ be the fixed-point of this hierarchy. This will serve as our grounding hierarchy.

The set of grounded truths is extracted as usual: let $\Gamma_0 := \varnothing$, put $\Gamma_{\alpha+1} := \{\varphi \in \Sigma_{\alpha+1} | (\mathbb{N}, \Gamma_\alpha) \vDash \varphi\}$, and take unions at limit points. Let Γ be the fixed-point of this hierarchy. Then $(\mathbb{N}, \Sigma, \Gamma) \vDash$ G0 + G1 + G3-G8 + T1-T7, as is easily verified.

One might also start with $\Sigma_0 := \mathcal{L}_{PA} \cup \{\varphi | \mathsf{PAT} \vdash \varphi\}$, thus obtaining a model of G0-G8 + T1-T7. I assume that the models so obtained are proper subsets of the minimal Leitgeb fixed point, but I have no proof of this.

Ignoring for a moment the special Weak Kleene axiom G5^c, it is easily seen that the Leitgeb fixed points satisfy all axioms except those for conditional dependence. The Strong Kleene fixed points satisfy all axioms except G2, which states that all theorems of PAT are grounded. The supervaluational fixed points satisfy all of the axioms presented above:[4]

Definition 6.2.1. 1. LG := PAT + G0-G8 + T1-T7

2. WKG := PAT + G0-G1 + G3-G7 + G5^c + T1-T7

3. SKG := PAT + G0-G1 + G3-G12 + T1-T7

4. VFG := PAT + G0-G12 + T1-T7

We might also consider subsystems of the above theories that are obtained by dropping the compositional truth axioms. A theory which should be acceptable from a minimalist point of view is given by PAT + G1 + G8 + T7. G1 says that all T-free sentences are grounded, while G8 says that the result of attributing truth or falsity to sets of grounded sentences yields again a grounded sentence. T7 is just the relativized T-schema. Let us call this theory MG. It is easily seen that TB is a subtheory of MG. In fact, MG contains the theory UTB, i.e. the theory that extends PAT by all instances of the (strong) uniform T-schema

$$T\ulcorner\varphi(\dot{t_1}, \ldots, \dot{t_n})\urcorner \leftrightarrow \varphi(t_1^\circ, \ldots, t_n^\circ)$$

where φ is a T-free sentence. The acronym stands for 'uniform Tarski-biconditionals'. As we will see in the next section, MG actually interprets ϵ_0-many iterations of UTB. The following is obvious:

Theorem 6.2.2. *1. If $S = \mathcal{J}_L(S)$, then $(\mathbb{N}, \mathcal{J}_L(S)^+) \models$ LG.*

2. If $S = \mathcal{J}_{WK}(S)$, then $(\mathbb{N}, \mathcal{J}_{WK}(S)^+) \models$ WKG.

3. If $S = \mathcal{J}_{SK}(S)$, then $(\mathbb{N}, \mathcal{J}_{SK}(S)^+) \models$ SKG.

4. If $S = \mathcal{J}_{FV}(S)$, then $(\mathbb{N}, \mathcal{J}_{FV}(S)^+) \models$ VFG.

Under the Leitgeb and the supervaluational scheme, we start with the same initial set of grounded sentences —the arithmetical sentences plus all all theorems of PAT. But the supervaluational hierarchy grows more in width —it satisfies the axioms for conditional dependence, whereas Leitgeb's does not. This is why $(\mathbb{N}, \mathcal{J}_L^\infty(\varnothing)^+)$ does not satisfy the claim that Modus Ponens preserves truth (cf. section 5.2). The

[4] In the above definition, we assume that PAT now comprises induction axioms for the full language (including the new predicate G).

minimal Strong Kleene fixed point, on the other hand, satisfies the same growth axioms as the supervaluational one, but it starts with a smaller initial set—it only satisfies G1 but not G2; this is why the Strong Kleene fixed point does not satisfy the global reflection principle for PAT.

6.3. Proof-theoretic analysis

$(\mathbb{N}, \mathcal{J}^{\infty}_{SK}(\varnothing)^{+})$ is usually axiomatized by the Kripke-Feferman system KF, which has the same proof-theoretic strength as the system of Ramified Analysis up to ϵ_0, RA_{ϵ_0}, while $(\mathbb{N}, \mathcal{J}^{\infty}_{FV}(\varnothing)^{+})$ is usually axiomatized by Cantini's VF, which has the same proof-theoretic strength as the system ID_1 of elementary inductive definitions. What can be said about the proof-theoretic strength of the systems LG, WKG, SKG, VFG and MG?

Proposition 6.3.1. *Assume* PAT *as background theory. Then:*

1. $G1 + T7 \vdash \forall t_1 \dots t_n(T\ulcorner\varphi(\underset{.}{t_1}, \dots, \underset{.}{t_n})\urcorner \leftrightarrow \varphi(t_1^{\circ}, \dots, t_n^{\circ}))$, *for all* $\varphi \in \mathcal{L}_{PA}$.

2. $G0 + T7 \vdash \forall t_1 \dots \forall t_n(T\ulcorner\varphi(\underset{.}{t_1}, \dots, \underset{.}{t_n})\urcorner \rightarrow \varphi(t_1^{\circ}, \dots, t_n^{\circ}))$, *for all* $\varphi \in \mathcal{L}_T$.

3. $G0 + G4 + T3 \vdash \forall x(Sent_T(x) \rightarrow (T\underset{.}{\neg}x \rightarrow \neg Tx))$.

4. $G0 + G4 + T3 \vdash \forall x(Sent_T(x) \rightarrow (T\underset{.}{\neg}\underset{.}{\neg}x \leftrightarrow Tx))$.

5. $G0 + G7 + T6 \vdash \forall x \forall v(Sent_T(\underset{.}{\forall} vx) \rightarrow (\forall t T(x(t/v)) \rightarrow T(\underset{.}{\forall} vx)))$.

6. $G0 + T6 \vdash \forall x \forall v(Sent_T(\underset{.}{\forall} vx) \rightarrow (T(\underset{.}{\forall} vx) \rightarrow \forall t T(x(t/v))))$.

7. $G0 + G11 + T3 + T5 \vdash \forall x \forall y(Sent_T(x \underset{.}{\rightarrow} y) \rightarrow (T(x \underset{.}{\rightarrow} y) \rightarrow (Tx \rightarrow Ty)))$.

8. $G0 + G12 + T4 \vdash \forall x \forall y(Sent_T(x \underset{.}{\wedge} y) \rightarrow (T(x \underset{.}{\wedge} y) \leftrightarrow Tx \wedge Ty))$.

9. $G0 + G3 + T2 \vdash \forall t(T\underset{.}{T}t \leftrightarrow Tt^{\circ})$.

10. $G0 + G4 + G9 + G11 + T3 + T5 + T7 \vdash T\ulcorner\varphi\urcorner \vee T\ulcorner\neg\varphi\urcorner \rightarrow ((\varphi \rightarrow T\ulcorner\psi\urcorner) \leftrightarrow (T\ulcorner\varphi \rightarrow \psi\urcorner))$.

11. PAT $\vdash \varphi \Rightarrow$ *G2 + T7* $\vdash T\ulcorner\varphi\urcorner$.

Proof. Straightforward. For example, in order to prove (7), assume that Tx and $T(x \underset{.}{\rightarrow} y)$, i.e. $T(\underset{.}{\neg}x \underset{.}{\vee} y)$ holds. Tx and (4) imply $T\underset{.}{\neg}\underset{.}{\neg}x$. An application of grounding axiom G11 yields Gy. By axiom G4 we also get that $G\underset{.}{\neg}x$. Thus by our initial assumption and truth axiom T5 we have $T\underset{.}{\neg}x \vee Ty$. But from $T\underset{.}{\neg}\underset{.}{\neg}x$ and axiom T3 we get $\neg T\underset{.}{\neg}x$. Thus Ty must hold, as desired. □

Theorem 6.3.2. *PAT + G0-G2 + G4 + G7 + G9 + G11 + T1-T7 relatively interprets ID_1^{Acc}.*

Proof. Cantini [11] has shown that properties (1), (2), (5), (7), (10) and (11) of Proposition 6.3.1 suffice to establish the desired result. See also Halbach [34, chap. 26]. □

Since ID_1^{Acc} and ID_1 prove the same arithmetical theorems, we obtain:

Corollary 6.3.3. *VFG proves all arithmetical theorems of ID_1.*

Next, we will show that that the systems LG and SKG are at least as strong as the systems of ramified analysis up to ϵ_0. In order to do so, we show that both LG and SKG are able to define all truth-predicates of the system of Ramified Truth RT_{ϵ_0}. We first define sublanguages L_α of $\mathcal{L}_T$ for each $\alpha \prec \epsilon_0$. Those sublanguages of $\mathcal{L}_T$ can be seen as a translation of the Tarskian hierarchy of truth. This translation goes back to an idea of Kripke [53, p. 710], but seems to have made its first formally precise appearance in a paper by Halbach [35].

Definition 6.3.4. The sublanguages L_α of $\mathcal{L}_T$ are defined by recursion over the ordinals up to ϵ_0. L_0 is just $\mathcal{L}_{PA}$. For $0 \prec \alpha \prec \epsilon_0$, φ is a formula of the language L_α iff there are $\beta_1, \ldots \beta_n \prec \alpha$ such that every occurrence of a subformula Tt of φ occurs in the context $Sent(\overline{\beta_i}, t) \wedge Tt$ for some $0 < i \leqslant n$, where $Sent(\overline{\beta_i}, x)$ represents the set of L_{β_i}-sentences.

We can define 'x is a sentence of L_α' as follows, using Kleene's recursion theorem (where $OT(x)$ expresses that x is an ordinal term):

$$Sent(\alpha, x) \leftrightarrow [OT(\alpha) \wedge \exists \sigma, \tau < x(lh(\sigma) = lh(\tau) \wedge Rel(x, \sigma) \wedge$$

$$\wedge \forall u < lh(\tau)(OT(\tau(u)) \wedge \tau(u) \prec \alpha \wedge \sigma(u) = \ulcorner Sent(\tau(\dot{u}), v_0)\urcorner))]$$

Using transfinite induction, one can then show that the languages L_α for $\alpha \prec \epsilon_0$ are provably grounded:

Proposition 6.3.5. *For all $\delta \prec \epsilon_0$,*

$$\mathsf{PAT} + G1 + G8 + T7 \vdash \forall \zeta \prec \overline{\delta} \forall x(Sent(\zeta, x) \rightarrow G(x)).$$

Proof. Let $\varphi(v)$ be the formula $\forall x(Sent(v, x) \rightarrow G(x))$. PAT proves transfinite induction for every $\delta \prec \epsilon_0$, i.e. for all $\delta \prec \epsilon_0$ PAT proves:

$$\forall \alpha(\forall \beta \prec \alpha \varphi(\beta) \rightarrow \varphi(\alpha)) \rightarrow \forall \zeta \prec \overline{\delta} \varphi(\zeta).$$

So assume

$$\forall \beta \prec \alpha \forall x(Sent(\beta, x) \rightarrow G(x)). \;\; (I.H.)$$

Then it suffices to show that

$$\forall x(Sent(\alpha, x) \to G(x)).$$

Therefore let x be given and assume $Sent(\alpha, x)$. Then PA proves

$$OT(\alpha) \wedge \exists \sigma, \tau < x(lh(\sigma) = lh(\tau) \wedge Rel(x, \sigma) \wedge$$

$$\wedge \forall u < lh(\tau)(OT(\tau(u)) \wedge \tau(u) \prec \alpha \wedge \sigma(u) = \ulcorner Sent(\tau(\dot{u}), v_0)\urcorner)).$$

Let $\sigma, \tau < x$ and $u < lh(\tau) = lh(\sigma)$ be as above. Because the formula $Sent(\tau(u), v_0)$ is arithmetical and PAT + G1 + T7 proves the uniform T-biconditionals for all $\mathcal{L}_{PA}$-formulae, we get

$$\forall u \forall v_0 (T\ulcorner Sent(\tau(\dot{u}), \dot{v_0})\urcorner \leftrightarrow Sent(\tau(u), v_0)). \tag{6.1}$$

Since $\tau(u) \prec \alpha$, (I.H.) yields

$$\forall z(Sent(\tau(u), z) \to G(z)). \tag{6.2}$$

Because $\sigma(u) = \ulcorner Sent(\tau(\dot{u}), v_0)\urcorner$, (6.1) and (6.2) yield

$$\forall z(T\underset{\cdot}{s}(\sigma(u), z) \to G(z)).$$

Since this holds for all $u < lh(\sigma)$, Axiom G8 yields $G(x)$. □

In what follows, we write $\varphi_\alpha(t)$ for the formula $Sent(\overline{\alpha}, t) \wedge Tt$.

Proposition 6.3.6. *For all $\alpha \prec \epsilon_0$,* PAT + *G1* + *G3-G8* + *T1-T7 proves:*

1. $\forall s \forall t(\varphi_\alpha(s \underset{\cdot}{=} t) \leftrightarrow s^\circ = t^\circ)$
2. $\forall x(Sent(\overline{\alpha}, x) \to (\varphi_\alpha(\underset{\cdot}{\neg} x) \leftrightarrow \neg\varphi_\alpha(x)))$
3. $\forall x \forall y(Sent(\overline{\alpha}, x \underset{\cdot}{\wedge} y) \to (\varphi_\alpha(x \underset{\cdot}{\wedge} y) \leftrightarrow \varphi_\alpha(x) \wedge \varphi_\alpha(y)))$
4. $\forall x \forall y(Sent(\overline{\alpha}, x \underset{\cdot}{\vee} y) \to (\varphi_\alpha(x \underset{\cdot}{\vee} y) \leftrightarrow \varphi_\alpha(x) \vee \varphi_\alpha(y)))$
5. $\forall x \forall v(Sent(\overline{\alpha}, \underset{\cdot}{\forall} vx) \to (\varphi_\alpha(\underset{\cdot}{\forall} vx) \leftrightarrow \forall t \varphi_\alpha(x(t/v))))$
6. $\forall t(Sent(\overline{\beta}, t^\circ) \to (\varphi_\alpha(\underset{\cdot}{\varphi_\beta}(t)) \leftrightarrow \varphi_\beta(t^\circ)))$ *for* $\beta \prec \alpha$
7. $\forall t \forall \beta \prec \overline{\alpha}(Sent(\overline{\beta}, t^\circ) \to (\varphi_\alpha(\underset{\cdot}{\varphi_\beta}(t)) \leftrightarrow \varphi_\alpha(t^\circ)))$

Proof. (1) is just a restriction of T1. (2) follows from T3 and the fact that every sentence of L_α is grounded by Lemma 6.3.5. For (3), just observe that if $x \wedge y$ is a sentence of L_α, then both x and y are also sentences of L_α. Thus, the claim follows by Lemma 6.3.5 and T4. (4) and (5) are proved in the same manner. For (6), use T7 and Lemma 6.3.5; for (7), use T2 and Lemma 6.3.5. □

An application of the recursion theorem then shows that both LG and SKG relatively interpret RT_{ϵ_0}. (Proposition 6.3.6 alone does not establish the truth definability of the Tarskian hierarchy, because the predicates φ_α apply to syntactically different sentences.)

Proposition 6.3.7. *The recursion theorem for primitive recursive functions yields the existence of a primitive recursive translation function* $\tau : \mathcal{L}_T^{\epsilon_0} \to \mathcal{L}_T$ *such that:*

$$\tau(\psi) = \begin{cases} s = t, & \textit{if } \psi := s = t \\ \varphi_\alpha(\underset{\cdot}{\tau}(t)) & \textit{if } \psi := T_\alpha t \\ \neg\tau(\chi) & \textit{if } \psi := \neg\chi \\ \tau(\chi) \wedge \tau(\delta) & \textit{if } \psi := \chi \wedge \delta \\ \forall x \tau(\chi) & \textit{if } \psi := \forall x \chi \end{cases}$$

where $\underset{\cdot}{\tau}$ *is a function symbol for* τ *in* $\mathcal{L}_T$*.*

Using Proposition 6.3.6, one can then show:

Proposition 6.3.8. *For all* $\alpha \prec \epsilon_0$ *and* $\varphi \in \mathcal{L}_T^{\epsilon_0}$*:*

$$\textit{If } \mathsf{RT}_\alpha \vdash \varphi, \textit{ then } \mathsf{PAT} + \mathit{G1} + \mathit{G3}\text{–}\mathit{G8} + \mathit{T1}\text{–}\mathit{T7} \vdash \tau(\varphi).$$

As an immediate consequence we get:

Theorem 6.3.9. *Both* LG *and* SKG *define all truth predicates of* RT_{ϵ_0}.

Thus, SKG is at least as strong as KF, and VFG is as least as strong as VF. Hence, nothing is lost (in proof-theoretic strength) when we pass from KF to SKG or from VF to VFG. The system LG for Leitgeb's theory proves all arithmetical sentences of KF. Let us next investigate the system WKG. We need the following result by Fujimoto.

Theorem 6.3.10 (Fujimoto [29]). *Let* $\mathcal{S} \supseteq \mathsf{PA}$ *be an* $\mathcal{L}_T$*-theory that derives the following, for some formula* $D(x)$*.*

1. $D(x) \leftrightarrow (Tx \leftrightarrow \neg T \underset{\cdot}{\neg} x)$

2. $\forall s \forall t (T(s \underset{\cdot}{=} t) \leftrightarrow s^\circ = t^\circ)$

3. $\forall s \forall t (T(s \underset{\cdot}{\neq} t) \leftrightarrow s^\circ \neq t^\circ)$

4. $(D(x) \wedge D(y)) \to (D(\underset{\cdot}{\neg} x) \wedge D(x \underset{\cdot}{\wedge} y) \wedge D(x \underset{\cdot}{\to} y))$

5. $D(x \underset{\cdot}{\vee} y) \to (T(x \underset{\cdot}{\vee} y) \leftrightarrow (Tx \vee Ty))$

6. *$D(x \underset{.}{\to} y) \to (T(x \underset{.}{\to} y) \leftrightarrow (Tx \to Ty))$*

7. *$D(\underset{.}{\forall} vx) \to (T(\underset{.}{\forall} vx) \leftrightarrow \forall t Tx(t/v))$*

8. *Transfinite induction in $\mathcal{L}_T$ up to (but not necessarily including) α.*

Then $\mathcal{S}$ defines all truth predicates of RT_α.

Theorem 6.3.11. WKG *defines all truth predicates of* RT_{ϵ_0}.

Proof. It suffices to show that WKG satisfies Theorem 6.3.10, taking the predicate $G(x)$ for $D(x)$ and ϵ_0 for α. By G0 and T3 it follows that $G(x)$ satisfies item 1. Item 2+3 follow from G1+T7. Item 4 is an immediate consequence of G4-G6. Item 5 follows from G5^c+T5. Item 6 follows from G5^c + T3 + T5. We give the proof in some detail. Notice that $T(x \underset{.}{\to} y)$ is defined as $T(\underset{.}{\neg} x \underset{.}{\vee} y)$.

1.	$G(\underset{.}{\neg} x \underset{.}{\vee} y)$	Premise 1
2.	$T(\underset{.}{\neg} x \underset{.}{\vee} y)$	Premise 2
3.	Tx	Premise 3
4.	$G\underset{.}{\neg} x \wedge Gy$	1, Axiom G5^c
5.	$\neg T \underset{.}{\neg} x$	3, 4, Axiom G4, T3
6.	$T\underset{.}{\neg} x \vee Ty$	2, 4, Axiom T5
7.	Ty	5, 6, Modus Ponens
8.	$G(\underset{.}{\neg} x \underset{.}{\vee} y) \to (T(\underset{.}{\neg} x \underset{.}{\vee} y) \to (Tx \to Ty))$	1-7

1.	$G(\underset{.}{\neg} x \underset{.}{\vee} y)$	Premise 1
2.	$Tx \to Ty$	Premise 2
3.	$\neg Tx \vee Ty$	2, definition →
4.	$G\underset{.}{\neg} x \wedge Gy$	1, Axiom G5^c
5.	$\neg Tx \leftrightarrow T\underset{.}{\neg} x$	4, Axiom T3
6.	$T\underset{.}{\neg} x \vee Ty$	3, 5, logic
7.	$T(\underset{.}{\neg} x \underset{.}{\vee} y)$	4, 6, Axiom T5
8.	$G(\underset{.}{\neg} x \underset{.}{\vee} y) \to ((Tx \to Ty) \to T(\underset{.}{\neg} x \underset{.}{\vee} y))$	1-7, logic

Item 7 follows from G0+T6. Item 8 is satisfied because PAT is a subtheory of WKG. □

Finally, let us have a look at the minimalist theory MG. Let IUTB_α be the result of iterating the typed disquotational theory UTB α-many times. More precisely, IUTB_α is the theory in the language of the Tarskian truth hierarchy $\mathcal{L}_T^{\alpha+1}$ that is obtained from PAT by adding typed uniform T-biconditionals

$$\forall t_1 \ldots \forall t_n (T_\alpha \ulcorner \varphi(\underset{.}{t_1} \ldots \underset{.}{t_n}) \urcorner \leftrightarrow \varphi(t_1^\circ \ldots t_n^\circ)$$

for every formula $\varphi \in \mathcal{L}_T^\alpha$. The acronym IUTB stands for 'iterated uniform Tarski-biconditionals'. Proposition 6.3.5 implies:

Proposition 6.3.12. *For every $\varphi \in L_\alpha$ with $\alpha \prec \epsilon_0$ we have:*

$$\mathsf{MG} \vdash \forall t_1 \dots \forall t_n (T\ulcorner \varphi(\underset{\cdot}{t_1} \dots \underset{\cdot}{t_n})\urcorner \leftrightarrow \varphi(t_1^\circ \dots t_n^\circ)$$

Corollary 6.3.13. MG *defines all truth predicates of* $\mathsf{IUTB}_{\epsilon_0}$.

I conjecture that MG is conservative over PAT, although I have no proof of this. However, by adding *uniform* reflection principles to MG, one can obtain a strong theory. Halbach [37] has shown that reflecting on UTB already yields the axioms of the typed compositional theory CT. One might conjecture that uniform reflection for MG yields the Tarskian hierarchy RT up to the ordinal ϵ_0.

6.4. Comparison

We conclude with some final remarks. All of the compositional systems introduced above prove the consistency axiom (Cons). It can be consistently added to KF, but is usually not a part of it, because it differs in character from the other axioms of KF.[5] Here we have it as a consequence. Furthermore, T-Out and the claim that Modus Ponens preserves truth are also consequences of our theories, but they are again no part of KF in its usual setting, even though they can be consistently added; in fact, they are consequences of KF + (Cons).

The systems LG and VFG prove the weak T-rule (property (11) of Proposition 6.3.1). As far as I can see, it is not possible to derive the stronger global reflection principle for PAT in these systems. It is possible to prove that all axioms of predicate logic are true, that all axioms of PA are true and that the usual inference rules of predicate logic are truth-preserving. However, the most straightforward proof that all instances of induction containing the truth predicate are true requires that the truth predicate is complete, i.e. that for all x, either Tx or $T\underset{\cdot}{\neg}x$. And this is not a theorem of LG or of VFG.[6] Of course, the reflection principle can consistently be added to both LG and VFG (it can also consistently be added to SKG, but the principle is not sound with respect to the Strong Kleene fixed points). I refrained from doing so because it differs in character from the other axioms.

Notice that VFG plus global reflection for PAT proves axioms V1−V8 of Cantini's VF. As we have already remarked, some authors complain that the axioms of VF seem somewhat unrelated and lack a common denominator. I hope the present axiomatization shows that it is possible to reformulate VF in a way that has the appearance of a theory of grounded truth.

The disquotational theory MG might be of interest to authors such as Horwich. In his paper [47], Horwich proposes to restrict the T-schema to sentences that are

[5] Cf. the discussion in Feferman [20], pp. 19−20.

[6] Notice that the completeness axiom is inconsistent with T-Out.

grounded. However, he explicitly rejects Kripke's approach because it "invokes Tarski-style compositional rules" (p. 82, fn 11). Horwich therefore suggests that the concept of grounding may be adapted in such a way that it squares with minimalism. He proposes to regard a sentence φ as grounded iff

> [φ or its negation $\neg\varphi$] is entailed either by the non-truth-theoretic facts, or by those facts together with whichever truth-theoretic facts are 'immediately' entailed by them (via the already legitimized instances of the equivalence schema), or ... and so on. [47, p. 81]

We might try to formalize this proposal as follows. Let $H_0 := Th(\mathbb{N})$ be the theory of the standard model of arithmetic, i.e. the set of arithmetical truths. Let $\Gamma_n := H_n + T \restriction H_n$, where $T \restriction H_n$ denotes the T-schema restricted to members of H_n. Then let

$$H_{n+1} := \{\varphi \in \mathcal{L}_T | \Gamma_n \vDash \varphi \text{ or } \Gamma_n \vDash \neg\varphi\}$$

where $\vDash$ is first-order consequence. Then we may let $H := H_\omega := \bigcup_n H_n$ be the set of grounded sentences (in the sense of Horwich).

Notice that H satisfies axioms G1, G2, G3, G4, G5 and G6, but not G7 or G8. So the problem with Horwich's proposal is that there will be many intuitively grounded sentences that won't count as grounded according to his definition. For example, while both $T\ulcorner\varphi\urcorner, T\ulcorner\neg\varphi\urcorner$ will be in H for every $\mathcal{L}_{PA}$-sentence φ, the sentence $\forall x(Sent_{PA}(x) \to Tx \vee T\underset{\cdot}{\neg}x)$ won't be in H, because such universal statements are not logically implied (in first-order logic) by their instances. Furthermore, iterating the above construction into the transfinite is of no use, because anything entailed by H_ω is already entailed by some H_n, by the compactness of first-order logic (notice that in the above definition, we would get an extensionally equivalent definition if we replace $\vDash$ by $\vdash$). But a notion of grounding according to which the statement 'All sentences of the base language have a definite truth-value' is not grounded is hardly convincing.

We have seen (section 4.2) that Horwich flirts with the idea of the ω-rule. So suppose that we would replace $\vDash$ by $\vdash_\omega$ in the above defintion. Then, first, we could iterate the hierarchy into the transfinite (i.e. the fixed point would lie well beyond ω) and second, axioms G7 and G8 would be satisfied too.

So it seems that MG—in fact, the extension of MG by axioms G2-G7—is a theory not involving "Tarski-style compositional rules" and minimalists might embrace it. In chapter 9, we will propose disquotational theories similar in spirit to MG that are deductively much stronger than any of the theories considered so far.

Part III.

Truth, Definability, and Comprehension

7. Truth-sets and second-order structures

We have seen in chapter 1.2 that truth (or satisfaction) and set-theoretic membership are closely related. Given the uniform T-biconditional

$$\forall x(T\ulcorner\varphi(\dot{x})\urcorner \leftrightarrow \varphi(x))$$

we can interpret the syntactic object $\ulcorner\varphi(x)\urcorner$ as the set $\{x|\varphi(x)\}$. The objective of this and the following chapters is to investigate this relationship in a more systematic manner. We will first show how to canonically associate, with any extension (interpretation) of the truth predicate (which we call a 'truth-set'), a structure (interpretation) for the language of second-order arithmetic. Second, we will give a translation of the language of second-order arithmetic into the language of truth. We will show that the translation of a second-order sentence is true in a truth-set if and only if the original sentence is true in the second-order structure associated with the truth-set. This correspondence can be used for quite a few interesting purposes. We will show that if $S = (S^+, S^-)$ is the *minimal* Kripke fixed point under an appropriate valuation scheme V, then S^+ is Π^1_1-hard and that $(\mathbb{N}, S^+)$ is a model of (the translation of) the impredicative theory ID_1. For the minimal fixed points under the Strong Kleene and the supervaluational scheme, these results have already been shown by Cantini (cf. [10], [11]). The main innovation here is that our proof also applies to Leitgeb's theory of truth. We also show that *any* Kripke fixed point of an appropriate valuation scheme satisfies the disquotational theory PUTB and that PUTB, when taken over logic alone, interprets Robinson arithmetic Q. In chapter 8 we show, using similar techniques, that the sets definable over standard model of the Tarskian hierarchy RT are exactly the hyperarithmetic sets. In chapter 9 we utilize the correspondence to establish the consistency of disquotational theories of truth that are obtained by translating comprehension axioms into T-biconditionals. These results also show that disquotational theories of truth can be much stronger than our best compositional theories of truth.

7.1. The Translation Lemma

Let us first introduce the language of second-order arithmetic.

Definition 7.1.1. 1. The language $\mathcal{L}_2$ of second-order arithmetic is obtained from $\mathcal{L}_{PA}$ by adding the binary relation symbol $\in$ plus *set variables* $X_0, X_1, X_2, \ldots$ (Let us call $v_0, v_1, \ldots$ *number* variables.) This gives us new formulae of the form $t \in X$ and $\forall X \varphi$. $\mathcal{L}_2$ is a two-sorted first-order language with usual (first-order) rules for both set and number quantifiers. A formula φ of $\mathcal{L}_2$ is called *arithmetical* if does not contain bound set variables. (Free set variables are allowed.)

2. Standard models for $\mathcal{L}_2$ have the form $(\mathbb{N}, \mathcal{M})$, where $\mathcal{M} \subseteq \wp(\omega)$ and the set variables X_i range over the elements of $\mathcal{M}$.

Recall that standard models of $\mathcal{L}_T$ have the form $(\mathbb{N}, S)$, where $\mathbb{N}$ interprets the arithmetical vocabulary and $S \subseteq \omega$ interprets the truth predicate T. Let us call S a *truth-set*. Any truth-set $S \subseteq \omega$ gives rise to a canonical second-order structure $(\mathbb{N}, \mathcal{M}_S)$ as follows:

Definition 7.1.2. Let $S \subseteq \omega$ and $\varphi \in Form^1_T$ (=an $\mathcal{L}_T$-formula with exactly one free variable).

1. $S_\varphi = \{n | \#\varphi(\overline{n}) \in S\} \subseteq \omega$
2. $\mathcal{M}_S = \{S_\varphi | \varphi \in Form^1_T\} \subseteq \wp(\omega)$
3. $\mathcal{M}^{tot}_S = \{S_\varphi | \varphi \in Form^1_T, \varphi\ S\text{-total}\} \subseteq \wp(\omega)$

Here, a formula $\varphi(x)$ is called S-total iff $(\mathbb{N}, S) \vDash \forall x(T\ulcorner\varphi(\dot{x})\urcorner \vee T\ulcorner\neg\varphi(\dot{x})\urcorner)$. As before, we occasionally identify expression with their codes. Accordingly, we also write S_k for S_φ, provided that $k = \#\varphi$.

$(\mathbb{N}, \mathcal{M}_S)$ and $(\mathbb{N}, \mathcal{M}^{tot}_S)$ are structures for the language of second-order arithmetic, $\mathcal{L}_2$. In the terminology of Cantini [10], [11], $\mathcal{M}_S$ is the *envelope* of S, and $\mathcal{M}^{tot}_S$ is the *section* of S. We note the following:

Proposition 7.1.3. *If $S = (S^+, S^-)$ is a Kripke fixed point, then $\mathcal{M}_{S^+}$ coincides with the collection of sets that are* weakly definable *in S, and $\mathcal{M}^{tot}_{S^+}$ coincides with the collection of sets that are* strongly definable *in S.*

Proof. If S is a Kripke fixed point, then

$$\begin{aligned}(S^+)_\varphi &= \{n | \#\varphi(\overline{n}) \in S^+\} \\ &= \{n | V(S)(T\ulcorner\varphi(\overline{n})\urcorner) = 1\} \\ &= \{n | V(S)(\varphi(\overline{n})) = 1\}\end{aligned}$$

□

Furthermore, we observe the following:

Proposition 7.1.4. *If $\mathcal{M}_S$ contains all Π^1_n-sets (Σ^1_n-sets), then S is Π^1_n-hard (Σ^1_n-hard).*

Proof. By definition, S is Π^1_n-hard iff for every Π^1_n-set P there is a recursive function such that $n \in P$ iff $f(n) \in S$. Let P be given. Then by assumption $P = S_\varphi \in \mathcal{M}_S$ for some $\varphi(x)$. Set $f(n) := \#\varphi(\overline{n})$. □

Consider the following translation function from the language $\mathcal{L}_2$ to the truth language $\mathcal{L}_T$.

Definition 7.1.5. The function $^* : \mathcal{L}_2 \to \mathcal{L}_T$ is defined as follows:
$v_i^* = v_{2i}$, $X_i^* = v_{2i+1}$
$\overline{0}^* = \overline{0}$, $f(t_1, \ldots, t_n)^* = f(t_1^*, \ldots, t_n^*)$
$(s = t)^* = (s^* = t^*)$, $(\neg\varphi)^* = \neg\varphi^*$, $(\varphi \wedge \psi)^* = \varphi^* \wedge \psi^*$
$(t \in X_i)^* = T\underset{\cdot}{s}(v_{2i+1}, t^*)$
$(\forall v_i \varphi)^* = \forall v_{2i} \varphi^*$
$(\forall X_i \varphi)^* = \forall v_{2i+1}(Fm^1_T(v_{2i+1}) \to \varphi^*)$

Here, the predicate $Fm^1_T(x)$ naturally represents the set of (codes of) formulae of $\mathcal{L}_T$ that contain exactly one free variable; the function symbol $\underset{\cdot}{s}$ represents the substitution function described in the introduction. On the above translation, the formula $t \in X$ is translated as

> The result of substituting t for the free variable in (the formula) X^* is true

Sometimes it is convenient to add set constants to $\mathcal{L}_2$. Given $\mathcal{M}_S$, we let $\overline{S_\varphi}$ denote the set S_φ. We expand our above translation function by letting

$$(t \in \overline{S_\varphi})^* = T\underset{\cdot}{s}(\ulcorner\varphi\urcorner, t^*)$$

If h is a variable assignment for $(\mathbb{N}, \mathcal{M}_S)$, define the assignment h^* for $(\mathbb{N}, S)$ by $h^*(v_{2i}) = h(v_i)$ and $h^*(v_{2i+1}) = \min\{k \mid S_k = h(X_i), k \in Form^1_T\}$.

Proposition 7.1.6. *Let h be an assignment for $(\mathbb{N}, \mathcal{M}_S)$. Then $t^{(\mathbb{N},\mathcal{M}_S),h} = t^{*(\mathbb{N},S),h^*}$ for all number terms t of $\mathcal{L}_2$.*

The following important proposition shows that the translation of a second-order sentence is true in a truth-set if and only if the original sentence is true in the second-order structure associated with that truth-set.

Proposition 7.1.7 (Translation Lemma)**.** *Let $S \subseteq \omega$, let $\varphi(\vec{y}, \vec{X}, \vec{\overline{S_\gamma}}) \in \mathcal{L}_2$, and let h be an assignment for $(\mathbb{N}, \mathcal{M}_S)$. Then:*

$$(\mathbb{N}, \mathcal{M}_S), h \vDash \varphi \Leftrightarrow (\mathbb{N}, S), h^* \vDash \varphi^*$$

Proof. By induction on the complexity of formulae. The case $s = t$ is trivial.

Consider $t \in X_i$, where t is any term. Let $t^{(\mathbb{N},\mathcal{M}_S),h} = n$ and $h(X_i) = A$. There's a k such that $k = \min\{m \mid S_m = A\}$. Then

$$\begin{aligned}(\mathbb{N},\mathcal{M}_S), h \vDash t \in X_i &\Leftrightarrow n \in A\\ &\Leftrightarrow n \in S_k\\ &\Leftrightarrow s(k,n) \in S\\ &\Leftrightarrow (\mathbb{N},S) \vDash T\underset{\cdot}{s}(\overline{k},\overline{n})\\ &\Leftrightarrow (\mathbb{N},S), h^* \vDash T\underset{\cdot}{s}(v_{2i+1}, t^*)\end{aligned}$$

Consider $t \in \overline{S_\gamma}$, where t is any term. Let $t^{(\mathbb{N},\mathcal{M}_S),h} = n$. Then

$$\begin{aligned}(\mathbb{N},\mathcal{M}_S), h \vDash t \in \overline{S_\gamma} &\Leftrightarrow n \in S_\gamma\\ &\Leftrightarrow s(\#\gamma,n) \in S\\ &\Leftrightarrow (\mathbb{N},S) \vDash T\underset{\cdot}{s}(\ulcorner\gamma\urcorner,\overline{n})\\ &\Leftrightarrow (\mathbb{N},S), h^* \vDash T\underset{\cdot}{s}(\ulcorner\gamma\urcorner, t^*)\end{aligned}$$

The cases $\neg\psi, \psi \wedge \chi$ and $\forall x\psi$ follow easily from the I.H.

Finally, consider $\forall X_i\psi$ and let $M = \{k|\forall m(S_m = S_k \to k \leqslant m)\}$.

$$\begin{aligned}(\mathbb{N},\mathcal{M}_S), h \vDash \forall X_i\psi &\Leftrightarrow \forall A \in \mathcal{M}_S : (\mathbb{N},\mathcal{M}_S), h(A : X_i) \vDash \psi && (7.1)\\ &\Leftrightarrow \forall k \in Form^1_T : (\mathbb{N},\mathcal{M}_S), h(S_k : X_i) \vDash \psi && (7.2)\\ &\Leftrightarrow \forall k \in M : (\mathbb{N},\mathcal{M}_S), h(S_k : X_i) \vDash \psi && (7.3)\\ &\Leftrightarrow \forall k \in M : (\mathbb{N},S), h^*(k : v_{2i+1}) \vDash \psi^* && (7.4)\\ &\Leftrightarrow \forall k \in Form^1_T : (\mathbb{N},S), h^*(k : v_{2i+1}) \vDash \psi^* && (7.5)\\ &\Leftrightarrow (\mathbb{N},S), h^* \vDash \forall v_{2i+1}(Fm^1_T(v_{2i+1}) \to \psi^*) && (7.6)\end{aligned}$$

The implication from (7.3) to (7.2) follows from the definition of M and the extensionality of sets. The equivalence between (7.3) and (7.4) is given by the inductive hypothesis, since $[h(S_k : X_i)]^* = h^*(k : v_{2i+1})$ for every minimal k. The step from (7.4) to (7.5) is justified because in translated formulae, such as ψ^*, v_{2i+1} occurs only in contexts of the form $T\underset{\cdot}{s}(v_{2i+1}, t)$. By the definition of the sets S_k, if $S_m = S_k$ then

$$(\mathbb{N},S), h' \vDash T\underset{\cdot}{s}(\overline{k},t) \leftrightarrow T\underset{\cdot}{s}(\overline{m},t),$$

for any term t and assignment h'. □

We briefly introduce a second translation function. The $\mathcal{L}_T$-predicate $tot(x)$ is defined as $Fm^1_T(x) \wedge \forall y(Tx(\dot{y}) \vee T\underset{\cdot}{\neg}x(\dot{y}))$.

Definition 7.1.8. The function $^{**} : \mathcal{L}_2 \to \mathcal{L}_T$ is defined as the function * except for the following clause:

$(\forall X_i \varphi)^{**} = \forall v_{2i+1}(tot(v_{2i+1}) \to \varphi^{**})$

If h is a variable assignment for $(\mathbb{N}, \mathcal{M}_S^{tot})$, define the assignment h^{**} for $(\mathbb{N}, S)$ by $h^{**}(v_{2i}) = h(v_i)$ and $h^{**}(v_{2i+1}) = \min\{k \mid S_k = h(X_i), k \in Form^1_T, k \text{ is S-total}\}$.

Proposition 7.1.9 (Translation Lemma II)**.** *Let $S \subseteq \omega$, let $\varphi(\vec{y}, \vec{X}, \overrightarrow{S_\gamma}) \in \mathcal{L}_2$, and let h be an assignment for $(\mathbb{N}, \mathcal{M}_S^{tot})$. Then:*

$$(\mathbb{N}, \mathcal{M}_S^{tot}), h \vDash \varphi \Leftrightarrow (\mathbb{N}, S), h^{**} \vDash \varphi^{**}$$

Proof. Similar to the proof of Proposition 7.1.7. □

7.2. Complexity of fixed-point theories

The Translation Lemmata can be used to establish lower bounds on the recursion-theoretic complexity of certain semantical theories of truth. We will first relate Kripke fixed points to inductive sets.

We briefly recall some concepts and results from Moschovakis [62]. Suppose that $\varphi(x, X)$ is an arithmetical $\mathcal{L}_2$-formula (with all free variables displayed) in which X occurs only positively. The operator $\Gamma_\varphi : \wp(\omega) \to \wp(\omega)$ given by φ is defined by

$$\Gamma_\varphi(S) = \{n | \mathbb{N} \vDash \varphi(\overline{n}, \overline{S})\}$$

This operator is monotone in the sense that, whenever $S \subseteq S'$, then $\Gamma(S) \subseteq \Gamma(S')$. Let $I^0_\varphi = \varnothing$, $I^{\alpha+1}_\varphi = \Gamma_\varphi(I^\alpha_\varphi)$, and $I^\gamma_\varphi = \bigcup_{\alpha<\gamma} I^\alpha_\varphi$, when γ is a limit ordinal. Let $I_\varphi := I^\kappa_\varphi$ where κ is least with $I^\kappa_\varphi = I^{\kappa+1}_\varphi$.

If P is a Π^1_1-set then there is an $\mathcal{L}_{PA}$-formula $\psi(x)$ such that for all n, $n \in P$ iff $\langle\langle\varnothing\rangle, n\rangle \in I_\varphi$, where $\varphi(x, X)$ is

$$Seq((x)_0) \wedge (\psi(x) \vee \forall t \langle (x)_0 \frown t, (x)_1 \rangle \in X)$$

Here, $Seq(u)$ expresses that u is the code of a (finite) sequence of natural numbers, $u \frown t$ denotes the concatenation of the sequence u with the sequence $\langle t \rangle$, $\langle\varnothing\rangle$ denotes the empty sequence, $\langle\rangle$ is some pairing function, and $(x)_i$ refers to the i-th argument of x.

Now, given an arithmetical $\mathcal{L}_2$-formula $\varphi(v_0, X_0)$ with exactly the displayed variables free and X_0 occurring positively, let $\iota\varphi := \ulcorner \varphi^*(v_0, \underset{\cdot}{s}^2_2(v_1, v_1)) \urcorner$ and $I_{\iota\varphi} := \underset{\cdot}{s}^2_2(\iota\varphi, \iota\varphi)$. Observe that

$$\begin{aligned} I_{\iota\varphi} &= \underset{\cdot}{s}^2_2(\iota\varphi, \iota\varphi) \\ &= \ulcorner \varphi^*(v_0, \underset{\cdot}{s}^2_2(\iota\varphi, \iota\varphi)) \urcorner \\ &= \ulcorner \varphi^*(v_0, I_{\iota\varphi}) \urcorner \end{aligned}$$

The definition of the term $I_{\iota\varphi}$ is due to Cantini [10]. Observe that $\varphi^*(v, I_{\iota\varphi})$ is

$$Seq((v)_0) \wedge (\psi(v) \vee \forall t T\underset{\cdot}{s}(I_{\iota\varphi}, \langle (v)_0 \frown t), (v)_1\rangle)$$

Definition 7.2.1. A valuation scheme V is *nice* iff the following conditions hold:

1. if $\psi \in \mathcal{L}_{PA}$ and $\mathbb{N} \vDash \psi$ then $V(S)(\psi \vee \varphi) = 1$
2. if $V(S)(\varphi) = 1$ and $\psi \in \mathcal{L}_{PA}$ then $V(S)(\psi \vee \varphi) = 1$
3. a conjunction is true under V if both conjuncts are true under V
4. if for all n, $f(n) \in S$, then $V(S)(\forall x T\underset{\cdot}{f}(x)) = 1$
5. an $\mathcal{L}_{PA}$-sentence is true (false) under V iff it is true (false) in the standard model
6. V is classically sound.

Recall that a valuation scheme is classically sound if every sentence that has a definite truth value in a partial model has the same truth value in its close-off. The Strong Kleene scheme, the Leitgeb scheme and all supervaluational schemes are nice. The Weak Kleene scheme, however, is not nice, as it does not satisfy property (1) and (2).

Theorem 7.2.2. *Assume that V is nice and that $\varphi(x, X)$ is as above. Then for all $\alpha \in ON$ we have:*

$$I^\alpha_\varphi = (\mathcal{J}^\alpha_V(\varnothing)^+)_{\varphi^*} := \{n | V(\mathcal{J}^\alpha_V(\varnothing))(T\ulcorner\varphi^*(\overline{n}, I_{\iota\varphi})\urcorner) = 1\}$$

Proof. By transfinite induction on α.

$\alpha = 0$: Since $\mathcal{J}^0_V(\varnothing) = \varnothing$, we get $(\mathcal{J}^\alpha_V(\varnothing)^+)_{\varphi^*} = \varnothing = I^0_\varphi$.

$\alpha = \beta + 1$: By I.H. $(\mathcal{J}^\beta_V(\varnothing)^+)_{\varphi^*} = I^\beta_\varphi$. Let $n \in (\mathcal{J}^{\beta+1}_V(\varnothing)^+)_{\varphi^*}$, that is,

$$n \in \{m | V(\mathcal{J}^{\beta+1}_V(\varnothing))(T\ulcorner\varphi^*(\overline{m}, I_{\iota\varphi})\urcorner) = 1\}$$

By definition of Kripke jump, $V(\mathcal{J}^\beta_V(\varnothing))(\varphi^*(\overline{n}, I_{\iota\varphi})) = 1$. By classical soundness of V, we have $(\mathbb{N}, \mathcal{J}^\beta_V(\varnothing)^+) \vDash \varphi^*(\overline{n}, I_{\iota\varphi})$. Then the I.H. and the Translation Lemma yield $\mathbb{N} \vDash \varphi(\overline{n}, \overline{I^\beta_\varphi})$, whence $n \in I^{\beta+1}_\varphi$.

For the other direction, assume that $\mathbb{N} \vDash \varphi(\overline{n}, \overline{I^\beta_\varphi})$, whence by I.H. and Translation Lemma $(\mathbb{N}, \mathcal{J}^\beta_V(\varnothing)^+) \vDash \varphi^*(\overline{n}, I_{\iota\varphi})$. This implies $\mathbb{N} \vDash Seq((\overline{n})_0)$, and by property (5) of a nice valuation this also holds under V. Furthermore,

$$(\mathbb{N}, \mathcal{J}^\beta_V(\varnothing)^+) \vDash \psi(\overline{n}) \vee \forall t T\underset{\cdot}{s}(I_{\iota\varphi}, \langle (\overline{n})_0 \frown t, (\overline{n})_1\rangle)$$

By property (3) of a nice valuation, it suffices to show that

$$V(\mathcal{J}_V^{\beta}(\varnothing))(\psi(\overline{n}) \vee \forall t T\underset{\cdot}{s}(I_{\iota\varphi}, \langle(\overline{u})_0 \frown t, (\overline{n})_1\rangle)) = 1$$

Now, either $(\mathbb{N}, \mathcal{J}_V^{\beta}(\varnothing)^+) \vDash \psi(\overline{n})$, so the result follows from property (1) of a nice valuation. On the other hand, if $(\mathbb{N}, \mathcal{J}_V^{\beta}(\varnothing)^+) \vDash \forall t T\underset{\cdot}{s}(I_{\iota\varphi}, \langle(\overline{u})_0 \frown t, (\overline{n})_1\rangle)$, then the claim follows from property (4) and (2) of a nice valuation.

α is a limit ordinal: by I.H. and definition we get:

$$(\mathcal{J}_V^{\alpha}(\varnothing)^+)_{\varphi^*} = \bigcup_{\beta<\alpha}\{n|V(\mathcal{J}_V^{\beta}(\varnothing))(T\ulcorner\varphi^*(\overline{n}, I_{\iota\varphi})\urcorner) = 1\} = \bigcup_{\beta<\alpha} I_{\varphi}^{\beta} = I_{\varphi}^{\alpha}$$

□

Corollary 7.2.3. *If V is nice and $S = (S^+, S^-) = \mathcal{J}_V^{\infty}(\varnothing)$ then*

1. $\mathcal{M}_{S^+} \supseteq \{P|P$ *is* $\Pi^1_1\}$
2. S^+ *is* $\Pi^1_1 -$ *hard*

Proof. If P is a Π^1_1-set then for all n, $n \in P$ iff $(\langle\varnothing\rangle, n) \in I_{\varphi}$, so (1) follows from the previous theorem.

(2) follows from (1) and Proposition 7.1.4. □

Notice the following:

Corollary 7.2.4. *If V is nice, $S = (S^+, S^-) = \mathcal{J}_V^{\infty}(\varnothing)$ and S^+ is itself Π^1_1, then S^+ is Π^1_1-complete and $\mathcal{M}_{S^+} = \{P|P$ is $\Pi^1_1\}$.*

Proof. It suffices to show that $\mathcal{M}_{S^+} \subseteq \{P|P$ is $\Pi^1_1\}$. This follows, under the assumptions, from the fact that every set weakly definable over S is elementary on S. □

The above results imply that the extension of the minimal fixed points under the Strong Kleene scheme, the Leitgeb scheme and all supervaluational schemes are Π^1_1-complete. For the minimal fixed points under the Strong Kleene and the supervaluational scheme, these results have already been shown by Cantini (cf. [10], [11]). The main innovation here is that our proof also applies to Leitgeb's theory of truth.

The above results imply that the close-off of minimal fixed points of nice valuations satisfy the (translation of the) axioms of the system ID_1, to which we now turn.

Definition 7.2.5. The language of ID_1 extends the language $\mathcal{L}_{PA}$ by a predicate constant $\overline{I_{\varphi}}$ for every *arithmetical* $\mathcal{L}_2$-formula $\varphi(v_0, X_0)$ (with exactly the displayed variables free) in which the free set variable X_0 occurs only *positively* (i.e. it does not appear in the scope of an odd number of negation signs). We may identify expressions of the form $\overline{I_{\varphi}}(t)$ with $t \in \overline{I_{\varphi}}$ and regard $\overline{I_{\varphi}}$ as a set constant. On the intended interpretation, the set constant $\overline{I_{\varphi}}$ is interpreted by the least fixed point generated (or the inductive relation defined) by the formula φ.

Definition 7.2.6. ID_1 is the theory in $\mathcal{L}_{ID_1}$ that contains in addition to the axioms of PA and full induction in $\mathcal{L}_{ID_1}$ all axioms of the form

$$\forall x(\varphi(x, I_\varphi) \to I_\varphi(x))$$

and

$$\forall x(\varphi(x, \psi) \to \psi(x)) \to \forall x(I_\varphi(x) \to \psi(x))$$

Here, $\varphi(x, \psi)$ is obtained from $\varphi(x, X)$ by replacing every occurrence of $t \in X$ by $\psi(t)$ and of $\neg(t \in X)$ by $\neg\psi(t)$. The system $\widehat{\mathsf{ID}}_1$ is the theory in $\mathcal{L}_{ID_1}$ that contains in addition to the axioms of PA and full induction in $\mathcal{L}_{ID_1}$ all axioms of the form

$$\forall x(\varphi(x, I_\varphi) \leftrightarrow I_\varphi(x))$$

In order to translate ID_1 into $\mathcal{L}_T$, expand the above translation function * by letting $(\overline{I_\varphi})^* = I_{\iota\varphi}$ and $(t \in \overline{I_\varphi})^* = T\underset{\cdot}{s}(I_{\iota\varphi}, t^*)$.

Corollary 7.2.7. *If V is nice and $S = (S^+, S^-) = \mathcal{J}_V^\infty(\varnothing)$, then* $(\mathbb{N}, S^+) \vDash (\mathsf{ID}_1)^*$.

Now we will relate the Kripke fixed points to the hyperarithmetical sets. A set P is hyperarithmetical iff both P and its complement are inductive on $\mathbb{N}$ (i.e. both P and its complement are Π^1_1). Call a valuation scheme V *negation-normal* iff either both $\varphi, \neg\varphi$ have definite opposed truth values or neither has a definite truth value.

Theorem 7.2.8. *If $S = (S^+, S^-) = \mathcal{J}_V^\infty(\varnothing)^+$ and V is nice and negation-normal, then*

$$\mathcal{M}_S^{tot} = \{X \subseteq \omega | X \text{ is } \Delta^1_1\} = HYP.$$

Proof. "$\subseteq$": Assume that φ is S-total. We have to show that S_φ is Δ^1_1. Under the assumptions, Corollary 7.2.4 implies that both S_φ and $S_{\neg\varphi}$ are Π^1_1. Now by negation-normality, $\neg\varphi$ is S-total too, therefore $S_{\neg\varphi}$ is the complement of S_φ, which means that S_φ must be Σ^1_1. Thus S_φ is both Σ^1_1 and Π^1_1, hence it is Δ^1_1.

"$\supseteq$": If Y is hyperarithmetical then

$$\overline{n} \in \overline{Y} \leftrightarrow \varphi(\langle\langle\varnothing\rangle, \overline{n}\rangle, \overline{I_\varphi})$$

for some φ and

$$\begin{aligned} \overline{n} \notin \overline{Y} &\leftrightarrow \chi(\overline{m}, \overline{I_\psi}) \\ &\leftrightarrow \neg\varphi(\langle\langle\varnothing\rangle, \overline{n}\rangle, \overline{I_\varphi}) \end{aligned}$$

for some m and χ. This implies that $\varphi^*(\langle\langle\varnothing\rangle, x\rangle, I_{\iota\varphi})$ is S-total, and the latter defines Y in S. □

Definition 7.2.9. $\Delta^1_1 - \mathsf{CA}_0$ is the theory in $\mathcal{L}_2$ that contains in addition to the axioms of PA all comprehension axioms

$$\forall\vec{Y}\forall\vec{y}\forall x(\varphi(x,\vec{y},\vec{Y}) \leftrightarrow \psi(x,\vec{y},\vec{Y})) \to \forall\vec{Y}\forall\vec{y}\exists X\forall x(x \in X \leftrightarrow \varphi(x,\vec{y},\vec{Y})),$$

where $\varphi(x,\vec{y},\vec{Y}) \in \mathcal{L}_2$ is a Π^1_1-formula and $\psi(x,\vec{y},\vec{Y}) \in \mathcal{L}_2$ is a Σ^1_1-formula, and the induction axiom

$$\forall X(0 \in X \land \forall x(x \in X \to x + \overline{1} \in X) \to \forall x(x \in X)).$$

The minimal ω-model of the system $\Delta^1_1 - \mathsf{CA}_0$ is the structure $(\mathbb{N}, HYP)$.

Let S^{tot} consist of the codes of those sentences $\varphi(\overline{n})$ such that $\varphi(x)$ is S-total and $\#\varphi(\overline{n}) \in S$. Notice that $\mathcal{M}^{tot}_S = \mathcal{M}_{S^{tot}}$.

Corollary 7.2.10. *Let S be the extension of the truth predicate in the minimal fixed-point under the Weak Kleene, the Strong Kleene, the FV supervaluations or the Leitgeb valuation scheme. Then*

1. $(\mathbb{N}, S^{tot}) \vDash (\Delta^1_1 - \mathsf{CA}_0)^*$.
2. $(\mathbb{N}, S) \vDash (\Delta^1_1 - \mathsf{CA}_0)^{**}$.

Proof. Ad 1. Since $\mathcal{M}^{tot}_S = HYP$ by Theorem 7.2.8 and $(\mathbb{N}, HYP) \vDash \Delta^1_1 - \mathsf{CA}_0$, the claim follows from the Translation Lemma 7.1.7.

Ad 2. Since $\mathcal{M}^{tot}_S = HYP$ by Theorem 7.2.8 and $(\mathbb{N}, HYP) \vDash \Delta^1_1 - \mathsf{CA}_0$, the claim follows from the Translation Lemma II 7.1.9. □

The correspondence between truth-sets and second-order models that we have established in this chapter seems to me to be a good way to measure the amount of second-order quantification that a semantic theory of truth is able to mimick. The theorems of this section indicate then, I think, a certain lower bound on the proof-theoretic strength that we should expect from a good axiomatization of the minimal Kripke fixed points. For example, a semantic theory that encodes all inductive sets should be axiomatized by a theory that formalizes the theory of inductive definitions, ID_1. Looking back at the systems introduced so far, only VF and VFG meet this requirement, while KF, SKG, LG etc. fall short of our expectations.[1] Burgess has given a variant of KF, called KFB (the acronym stands for 'Kripke-Feferman-Burgess') that does have the same strength as ID_1. The additional strength over KF is obtained by adding a minimality axiom scheme which basically says that if $\varphi(x)$ satisfies the KF axioms, then all true sentences fall under the extension of $\varphi(x)$ (cf. Halbach [38, chap. 17] for details). Another way to strengthen KF, which has

[1] But note that KF is not intended as an axiomatization of the minimal SK fixed point but of all fixed points; not all of them satisfy ID_1.

already been pointed out by Cantini [10, p. 105], is to add the translation of the second axiom scheme of ID_1 to KF, that is

$$\mathsf{KF} + (\forall x(\varphi(x, \psi) \to \psi(x)) \to \forall x(I_\varphi(x) \to \psi(x)))^+ \vdash (\mathsf{ID}_1)^+$$

Theorem 7.2.7 shows that the resulting system is still sound with respect to the minimal Strong Kleene fixed point.

7.3. Positive disquotation

Before closing this chapter, we briefly mention some results on positive disquotation.

Definition 7.3.1. PUTB is the theory in $\mathcal{L}_T$ that extends PAT by all sentences of the form

$$\forall t_1 \ldots \forall t_n (T\ulcorner\varphi(\underset{\cdot}{t}_1, \ldots, \underset{\cdot}{t}_n)\urcorner \leftrightarrow \varphi(t_1^\circ, \ldots, t_n^\circ))$$

where φ is T-positive.

Cantini has shown that PUTB proves the existence of fixed points for elementary positive operators of $\mathcal{L}_{PA}$.

Theorem 7.3.2 (Cantini [10])**.** $\mathsf{PUTB} \vdash (\widehat{\mathsf{ID}}_1)^*$.

Proof. Since φ is X-positive, φ^* is T-positive, so the following is an axiom of PUTB:

$$\forall t_1 \forall t_2 (T\ulcorner\varphi^*(\underset{\cdot}{t}_1, \underset{\cdot}{t}_2)\urcorner \leftrightarrow \varphi^*(t_1^\circ, t_2^\circ))$$

Letting $t_2 = I_{\iota\varphi}$ we get

$$\forall t_1 (T\ulcorner\varphi^*(\underset{\cdot}{t}_1, I_{\iota\varphi})\urcorner \leftrightarrow \varphi^*(t_1^\circ, I_{\iota\varphi}))$$

which implies

$$\forall v_0 (T\ulcorner\varphi^*(\dot{v}_0, I_{\iota\varphi})\urcorner \leftrightarrow \varphi^*(v_0, I_{\iota\varphi}))$$

The latter is short for

$$\forall v_0 (T\underset{\cdot}{s}(\ulcorner\varphi^*(v_0, I_{\iota\varphi})\urcorner, v_0) \leftrightarrow \varphi^*(v_0, I_{\iota\varphi}))$$

Since $I_{\iota\varphi} = \ulcorner\varphi^*(v_0, I_{\iota\varphi})\urcorner$, substitution of identicals yields

$$\forall v_0 (T\underset{\cdot}{s}(I_{\iota\varphi}, v_0) \leftrightarrow \varphi^*(v_0, I_{\iota\varphi}))$$

which is the translation of $\forall v_0(v_0 \in I_\varphi \leftrightarrow \varphi(v_0, I_\varphi))$. □

We have already seen that the close-offs of minimal fixed points of nice valuations satisfy $(\mathsf{ID}_1)^*$ and therefore $(\widehat{\mathsf{ID}}_1)^*$. But do all of them also satisfy PUTB? If the valuation schemes are 'strong' in the sense below, the answer is 'yes'. In fact, this applies not only to the minimal fixed points but to all of them.

Definition 7.3.3. A valuation scheme V is *strong* iff the following conditions hold:

1. a T-free statement is true (false) under V iff it is true (false) in the standard model $\mathbb{N}$
2. if φ is true under V, then $\varphi \vee \psi$ is true under V
3. if for all n, $\varphi(\overline{n})$ true under V, then $\forall x \varphi$ is true under V
4. If $t^{\mathbb{N}} \in S^+$ then $V(S)(Tt) = 1$.
5. V is classically sound.

Observe that every strong valuation scheme is nice, but not necessarily the other way round. The Strong Kleene and *all* supervaluational schemes are nice. The Weak Kleene scheme and the Leitgeb scheme are not nice, because both violate condition (2).

Proposition 7.3.4. *Suppose that V is strong and $S = (S^+, S^-)$. Let φ be a T-positive formula (i.e. T does not occur in the scope of an odd number of negation signs). Then*

$$(\mathbb{N}, S) \vDash_V \varphi \Leftrightarrow (\mathbb{N}, S^+) \vDash \varphi$$

Here, the expression $(\mathbb{N}, S) \vDash_V \varphi$ is short for $V(S)(\varphi) = 1$. Notice that $(\mathbb{N}, S) \nvDash_V \varphi$ means that either $V(S)(\varphi) = 0$ or $V(S)(\varphi) = \frac{1}{2}$.

Proof. The left-to-right direction follows from the classical soundness of V. The right-to-left direction is proved by induction on the build-up of the T-positive formula φ. If φ is of the form $s = t$ or $s \neq t$, this follows from property (1) of a strong valuation scheme. If φ is of the form Tt, this follows from property (4) of a strong valuation. If φ is a disjunction, the claim follows from the induction hypothesis and property (2) of a strong valuation, and similarly for a quantified statement, using property (3). (Since φ is T-positive, the negation case does not occur.) □

Theorem 7.3.5. *If V is strong and $S = (S^+, S^-)$ is any fixed point of $\mathcal{J}_V$, then* $(\mathbb{N}, S^+) \vDash \mathsf{PUTB}$.

Proof. Let φ be a T-positive sentence and assume that $(\mathbb{N}, S^+) \vDash \varphi$. Then Proposition 7.3.4 shows that $(\mathbb{N}, S) \vDash_V \varphi$, which implies by the fixed point property that $(\mathbb{N}, S) \vDash_V T\ulcorner\varphi\urcorner$. So by the classical soundness of V we get $(\mathbb{N}, S^+) \vDash T\ulcorner\varphi\urcorner$. Now assume that $(\mathbb{N}, S^+) \nvDash \varphi$. Since V is classically sound, φ has value 0 or $\frac{1}{2}$ in the partial model S and consequently, $\varphi \notin S^+$. So $(\mathbb{N}, S^+) \vDash \neg T\ulcorner\varphi\urcorner$. So $(\mathbb{N}, S^+) \vDash \varphi \leftrightarrow T\ulcorner\varphi\urcorner$ for all T-positive φ, whence by standardness $(\mathbb{N}, S^+) \vDash \mathsf{PUTB}$. □

Question 7.3.6. *If S is the minimal (any) fixed point of the Leitgeb valuation scheme V_L, do we have $(\mathbb{N}, S^+) \vDash \mathsf{PUTB}$?*

We close this section with an interesting 'ontological' result. Let us denote by PUTB^- the theory, formulated in $\mathcal{L}_T$, that contains as axioms all sentences of the form

$$\forall x_1 \dots \forall x_n (T\ulcorner\varphi(\dot{x_1}, \dots, \dot{x_n})\urcorner \leftrightarrow \varphi(x_1, \dots, x_n))$$

where φ is T-positive, plus first-order logic. No non-logical axiom of PA is an axiom of this theory. The next theorem shows that positive disquotation, on its own, forces the existence of infinitely many objects.

Theorem 7.3.7. PUTB^- *relatively interprets Robinson arithmetic* Q.

Proof. By a result of Montagna & Mancini [61], it suffices to show that PUTB^- interprets adjunctive set theory, AS, i.e. interprets the following two claims:

$$\exists y \forall x \neg(x \in y)$$

and

$$\forall z \forall w \exists y \forall x (x \in y \leftrightarrow x \in w \vee x = z)$$

We show that both claims are derivable by interpreting $x \in y$ as $T\underset{\cdot}{s}(y, x)$. Now, since $x \neq x$ is a T-positive formula, the following is an axiom of PUTB^-:

$$\forall x (T\ulcorner\dot{x} \neq \dot{x}\urcorner \leftrightarrow x \neq x)$$

Written out in full this is:

$$\forall x (T\underset{\cdot}{s}(\ulcorner x \neq x\urcorner, x) \leftrightarrow x \neq x)$$

from which we deduce in pure logic

$$\exists y \forall x \neg T\underset{\cdot}{s}(y, x)$$

For the second axiom of AS, note that the formula $T\underset{\cdot}{s}(w, x) \vee x = z$ is T-Positive, so we have as an axiom:

$$\forall z \forall w \forall x (T\ulcorner T\underset{\cdot}{s}(\dot{w}, \dot{x}) \vee \dot{x} = \dot{z}\urcorner \leftrightarrow T\underset{\cdot}{s}(w, x) \vee x = z)$$

Let $\varphi := T\underset{\cdot}{s}(w,x) \lor x = z$. So the axiom really says:

$$\forall z \forall w \forall x (T\underset{\cdot}{s}(\underset{\cdot}{s}{}_2^2(\underset{\cdot}{s}{}_3^3(\ulcorner\varphi\urcorner, z), w), x) \leftrightarrow T\underset{\cdot}{s}(w,x) \lor x = z)$$

Instantiating the quantifiers $\forall z \forall w$ to z, w, we get:

$$\forall x (T\underset{\cdot}{s}(\underset{\cdot}{s}{}_2^2(\underset{\cdot}{s}{}_3^3(\ulcorner\varphi\urcorner, z), w), x) \leftrightarrow T\underset{\cdot}{s}(w,x) \lor x = z)$$

By existential weakening,

$$\exists y \forall x (T\underset{\cdot}{s}(y,x) \leftrightarrow T\underset{\cdot}{s}(w,x) \lor x = z)$$

Now we re-introduce the universal quantifers and get

$$\forall z \forall w \exists y \forall x (T\underset{\cdot}{s}(y,x) \leftrightarrow T\underset{\cdot}{s}(w,x) \lor x = z)$$

as desired. □

The above theorem should bear some relevance on discussions about deflationism and conservativity. According to (some versions of) deflationism, truth is a ‘thin’ notion that does not contribute anything to our knowledge of the world. Horsten [44], Ketland [49] and Shapiro [84] have argued then that conservativness over the arithmetical base theory is essential to deflationism. Special attention (cf. Field [23]) has been paid to the role of mathematical induction, because most theories of truth are conservative over their base theory if induction is not allowed for formulae containing the truth predicate, while many truth theories prove new arithmetical statements if full induction is available. In particular, uniform positive disquotation, if formulated over PAT, relatively interprets the theory $\widehat{\mathsf{ID}}_1$, while, if formulated over PA, does not prove any new arithmetical statements (as Cantini [10] has shown). However, even if uniform positive disquotation conservatively extends PA (if full induction is dropped), this does not mean that positive disquotation is metaphysically ‘light’, being highly non-conservative over logic.

8. Hyperarithmetic sets and ramified truth

In this section we consider the Tarskian hierarchy up to the level ω_1^{CK}. The main goal of this section is to prove that the sets definable in that hierarchy are exactly the hyperarithmetical sets. This was first proved by Halbach [33]. We give slightly different proofs due to the present author, using the methods introduced in the last chapter.

Definition 8.0.8. The standard model of $\mathsf{RT}_{\omega_1^{CK}}$ is defined as follows. Let

$$R_0 = \{\#\varphi | \mathbb{N} \vDash \varphi, \varphi \in \mathcal{L}_{PA}\}$$

and let

$$R_{\alpha+1} = \{\#\varphi | (\mathbb{N}, (R_\beta)_{\beta \leqslant \alpha}) \vDash \varphi, \varphi \in \mathcal{L}_T^{\alpha+1}\}$$

Let R_γ be the union of the R_α for $\alpha < \gamma$ if γ is a limit ordinal.

In the above definition, $R_\beta \subseteq \omega$ interprets the truth predicate T_β. Similar to the preceeding chapter, the truth-set R_β can be seen as encoding a collection of sets of natural numbers as follows. The following definition is not restricted to the R_β but applies to any interpretation of the language of ramified truth that respects the type restrictions.

Definition 8.0.9. For any sequence $(S_\alpha)_{\alpha<\omega_1^{CK}}$ with $S_\alpha \subseteq \mathcal{L}_T^\alpha$ and $\varphi \in \mathcal{L}_T^\alpha$ let

$$(S_\alpha)_\varphi = \{n | \#\varphi(\overline{n}) \in S_\alpha\}$$

and

$$\mathcal{M}_{S_\alpha} = \{(S_\alpha)_\varphi | \varphi \in \mathcal{L}_T^\alpha\}$$

Then $\mathcal{M}_{S_\alpha} \subseteq \wp(\omega)$. Hence

$$(\mathbb{N}, (\mathcal{M}_{S_\alpha})_{\alpha<\omega_1^{CK}}) = (\mathbb{N}, \mathcal{M}_{S_0}, \mathcal{M}_{S_1}, \ldots, \mathcal{M}_{S_\alpha}, \mathcal{M}_{S_{\alpha+1}}, \ldots)$$

is a structure for $\mathcal{L}_2^{\omega_1^{CK}}$, the language of predicative (ramified) analysis.

The $R_\alpha-$hierarchy defined in 8.0.8 is monotone. Moreover:

Proposition 8.0.10. *Let $\alpha < \beta$ and $\varphi \in \mathcal{L}_T^\alpha$. Then the following holds for all $n \in \omega$:*

$$\#\varphi(\overline{n}) \in R_\beta \Leftrightarrow \#\varphi(\overline{n}) \in R_\alpha$$

Proof. This is an immediate consequence of Definition 8.0.8. □

Corollary 8.0.11. *Let $\varphi \in \mathcal{L}_T^\alpha$. Then for all $\beta \geqslant \alpha$:*

$$(R_\alpha)_\varphi = (R_\beta)_\varphi$$

The goal of this section is to prove that the standard model of RT encodes exactly the hyperarithmetical sets, that is HYP $= \mathcal{M}_{R_{\omega_1^{CK}}}$. Kleene has shown that the hyperarithmetical sets are exactly those that are definable in the language of ramified analysis, $\mathcal{L}_2^{\omega_1^{CK}}$ (see below).

Definition 8.0.12. For $\alpha \leqslant \omega_1^{CK}$ let $\mathcal{L}_2^\alpha = \mathcal{L}_{PA} \cup \{X_i^\beta | \beta < \alpha, i \in \omega\}$, where X_i^β is a unary second-order predicate variable.

We first translate the formulae of ramified analysis into formulae of the language of the Tarskian hierarchy. As before, the main idea is to translate $t \in X^\beta$ as the result of substituting t for the free variable in the formula x is true at level β.

Definition 8.0.13. The translation function $^* : \mathcal{L}_2^{\omega_1^{CK}} \to \mathcal{L}_T^{\omega_1^{CK}}$ is defined as follows:
$(x_i)^* = x_{5^i}$, $(X_i^\alpha)^* = x_{3^n 5^i}$, where $n = \ulcorner \alpha \urcorner$
$0^* = 0$, $f(t_1, \dots, t_n)^* = f(t_1^*, \dots, t_n^*)$
$(s = t)^* = (s^* = t^*)$, $(\neg\varphi)^* = \neg\varphi^*$, $(\varphi \wedge \psi)^* = \varphi^* \wedge \psi^*$
$(t \in X_i^\alpha)^* = T_\alpha \underset{\cdot}{s}((X_i^\alpha)^*, t^*)$
$(\forall x \varphi)^* = \forall x^* \varphi^*$
$(\forall X_i^\alpha \varphi)^* = \forall (X_i^\alpha)^* (Fm_\alpha((X_i^\alpha)^*) \to \varphi^*)$

Here, $Fm_\alpha(x)$ expresses that x is a formula of $\mathcal{L}_T^\alpha$ with exactly one free variable.

If h is a variable assignment for $(\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha})$, define the assignment h^* for $(\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha})$ by $h^*(x_{5^i}) = h(x_i)$ and $h^*(x_{3^{\ulcorner \beta \urcorner} 5^i}) = \min\{\#\varphi \mid (S_\beta)_\varphi = h(X_i^\beta)\}$.

Proposition 8.0.14. *Let h be an assignment for $(\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha})$. Then $t^{(\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha}), h} = t^{*(\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}), h^*}$ for all number terms t of $\mathcal{L}_2^{\alpha+1}$.*

We have:

Proposition 8.0.15 (Translation Lemma III). *For all $\alpha < \omega_1^{CK}, \varphi \in \mathcal{L}_2^{\alpha+1}$:*

$$(\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha}), h \vDash \varphi \textit{ iff } (\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}), h^* \vDash \varphi^*$$

Proof. By induction on the complexity of formulae. The case $s = t$ follows from the previous proposition.

Consider $t \in X_i^\gamma$, where t is any term and $\gamma \leqslant \alpha$. Let n be the denotation of t under h in $\mathbb{N}$. and $h(X_i^\gamma) = (S_\gamma)_\psi$. There's a χ s.t. $\#(\chi) = \min\{\#\varphi \mid (S_\gamma)_\varphi = (S_\gamma)_\psi\}$. Then

$$\begin{aligned}
(\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha}), h \vDash t \in X_i^\gamma &\Leftrightarrow n \in (S_\gamma)_\psi \\
&\Leftrightarrow n \in (S_\gamma)_\chi \\
&\Leftrightarrow \#\chi(\bar{n}) \in S_\gamma \\
&\Leftrightarrow (\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}) \vDash T_\gamma \ulcorner \chi(\bar{n}) \urcorner \\
&\Leftrightarrow (\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}) \vDash T_\gamma \underset{\cdot}{s}(\ulcorner \chi \urcorner, \bar{n}) \\
&\Leftrightarrow (\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}), h^* \vDash T_\gamma \underset{\cdot}{s}(x_{3^{\ulcorner \gamma \urcorner} 5^i}, t^*)
\end{aligned}$$

The cases $\neg\psi, \psi \wedge \chi$ and $\forall x \psi$ follow easily from the I.H.

Finally, consider $\forall X_i^\gamma \psi$, where $\gamma \leqslant \alpha$, and let M_γ^ν be the set of formulae χ of $\mathcal{L}_T^\gamma$ whose code $\#(\chi)$ is the smallest generating $(S_\gamma)_\chi$, i.e. for any formula ζ of $\mathcal{L}_T^\gamma$, if $(S_\gamma)_\zeta = (S_\gamma)_\chi$ then $\#\chi \leqslant \#\zeta$.

$$\begin{aligned}
(\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha}), h \vDash \forall X_i^\gamma \psi &\Leftrightarrow \forall A \in \mathcal{M}_{S_\gamma} : (\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha}), h(A : X_i^\gamma) \vDash \psi && (8.1) \\
&\Leftrightarrow \forall \chi \in \mathcal{L}_T^\gamma : (\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha}), h((S_\gamma)_\chi : X_i^\gamma) \vDash \psi && (8.2) \\
&\Leftrightarrow \forall \chi \in M_\gamma^\nu : (\mathbb{N}, (\mathcal{M}_{S_\beta})_{\beta \leqslant \alpha}), h((S_\gamma)_\chi : X_i^\gamma) \vDash \psi && (8.3) \\
&\Leftrightarrow \forall \chi \in M_\gamma^\nu : (\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}), h^*(\ulcorner \chi \urcorner : x_{3^{\ulcorner \gamma \urcorner} 5^i}) \vDash \psi^* && (8.4) \\
&\Leftrightarrow \forall \chi \in \mathcal{L}_T^\gamma : (\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}), h^*(\ulcorner \chi \urcorner : x_{3^{\ulcorner \gamma \urcorner} 5^i}) \vDash \psi^* && (8.5) \\
&\Leftrightarrow (\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}), h^* \vDash \forall x_{3^{\ulcorner \gamma \urcorner} 5^i} (Fm_\gamma(x_{3^{\ulcorner \gamma \urcorner} 5^i}) \to \psi^*) && (8.6)
\end{aligned}$$

The implication from (8.3) to (8.2) is due to the extensionality of sets. The equivalence between (8.3) and (8.4) is given by the inductive hypothesis, since $[h((S_\gamma)_\chi : X_i^\gamma)]^* = h^*(\ulcorner \chi \urcorner : x_{3^{\ulcorner \gamma \urcorner} 5^i})$ for every minimal χ. The step from (8.4) to (8.5) is justified because in translated formulae, such as ψ^*, $x_{3^{\ulcorner \gamma \urcorner} 5^i}$ occurs only in the context of $T_\gamma \underset{\cdot}{s}(x_{3^{\ulcorner \gamma \urcorner} 5^i}, t)$. By the definition of the truth-sets $(S_\gamma)_\varphi$, if $\#(\varphi) \geqslant \#(\chi)$ and $(S_\gamma)_\varphi = (S_\gamma)_\chi$ then

$$(\mathbb{N}, (S_\beta)_{\beta \leqslant \alpha}), h^* \vDash T_\gamma \underset{\cdot}{s}(\ulcorner \varphi \urcorner, t) \leftrightarrow T_\gamma \underset{\cdot}{s}(\ulcorner \chi \urcorner, t)$$

for any term t and assignment h^*. □

Definition 8.0.16 (Ramified Analytical Sets)**.** The sets $\mathcal{RA}_\alpha$ are defined as follows. $\mathcal{RA}_0$ = the collection of arithmetically definable sets. $\mathcal{RA}_{\alpha+1}$ = the set of $X \subseteq \omega$ such that there is a $\varphi \in \mathcal{L}_2^{\alpha+1}$ and φ defines X in $(\mathbb{N}, (\mathcal{RA}_\beta)_{\beta \leqslant \alpha})$ (where we assume that the variables X^β take values in $\mathcal{RA}_\beta$ for $\beta \leqslant \alpha$). At limits we take unions.

Notice that this hierarchy is monotone. The hierarchy can be iterated beyond the ordinal ω_1^{CK}. Monotonicity will then yield an ordinal such that $\mathcal{RA}_\alpha = \mathcal{RA}_{\alpha+1}$. The least such α is called β_0 and is well beyond ω_1^{CK}. The hierarchy of ramified analytical sets is the second-order version of Gödel's constructible hierarchy. We have $\mathcal{RA}_\alpha = L_\alpha \cap \wp(\omega)$, where L_α refers to the α-th level of the constructible hierarchy. I mention here that $(\mathbb{N}, \mathcal{RA}_{\beta_0})$ is a model of classical analysis Z_2, i.e. of full second-order arithmetic. (More on this can be found in the Appendix.) This will be important in the next chapter. For now we are only interested in the following important result, which is due to Kleene.

Theorem 8.0.17 (Kleene [51])**.** *$\mathcal{RA}_{\omega_1^{CK}} = HYP$.*

In order to prove our main result, we need the following preliminary lemma.

Proposition 8.0.18. *Let $\alpha < \omega_1^{CK}$ and $\varphi \in \mathcal{L}_T^\alpha$. Assume that $\mathcal{RA}_\beta = \mathcal{M}_{R_\beta}$ for all $\beta < \alpha$. Then there is a $\psi \in \mathcal{L}_T^\alpha$ such that ψ is in the range of the translation function * and $(R_\alpha)_\varphi = (R_\alpha)_\psi$ (i.e. φ and ψ define the same set over the standard model of the Tarskian hierarchy).*

Proof. This is proved by an induction on the build-up of φ. The claim is trivial if φ is an equation. So let $\varphi(x) := T_\beta f(x)$, where $\beta < \alpha$ and $f(x)$ is some term of $\mathcal{L}_{PA}$. Halbach [34, p. 122] has shown that there is a formula $\chi(x) \in \mathcal{L}_2^\alpha$ such that

$$(\mathbb{N}, (\mathcal{R}_\gamma)_{\gamma \leqslant \beta}) \vDash T_\beta v \Leftrightarrow (\mathbb{N}, (\mathcal{RA}_\gamma)_{\gamma \leqslant \beta}) \vDash \chi(v)$$

(Roughly, $\chi(x)$ is a definition of the Tarskian truth predicate of level β.) Since by assumption $\mathcal{RA}_\gamma = \mathcal{M}_{R_\gamma}$ for all $\gamma < \alpha$, the translation lemma together with the above equivalence implies

$$(\mathbb{N}, (\mathcal{R}_\gamma)_{\gamma \leqslant \beta}) \vDash T_\beta f(x) \Leftrightarrow (\mathbb{N}, (\mathcal{R}_\gamma)_{\gamma \leqslant \beta}) \vDash \chi^*(f(x))$$

So $\chi^*(f(x))$, which is in the range of the translation function, defines the same set as $T_\beta f(x)$. If $\varphi(x)$ is not atomic, the claim follows easily from the induction hypothesis. □

Theorem 8.0.19. *$\mathcal{RA}_\alpha = \mathcal{M}_{R_\alpha}$ for all ordinals $\alpha < \omega_1^{CK}$.*

Proof. By transfinite induction. It is easily checked that $\mathcal{M}_{R_0}$ = the collection of arithmetically definable sets = $\mathcal{RA}_0$.

Successor case:

Let $X \in \mathcal{RA}_{\alpha+1}$. Thus $X = \{n | (\mathbb{N}, (\mathcal{RA}_\beta)_{\beta \leqslant \alpha}) \vDash \varphi(\overline{n})\}$ for some $\varphi(x) \in \mathcal{L}_2^{\alpha+1}$. Hence

$$\begin{aligned} n \in X &\Leftrightarrow (\mathbb{N}, (\mathcal{RA}_\beta)_{\beta \leqslant \alpha}) \vDash \varphi(\overline{n}) \\ &\Leftrightarrow (\mathbb{N}, (\mathcal{M}_{R_\beta})_{\beta \leqslant \alpha}) \vDash \varphi(\overline{n}) \text{ by I.H.} \\ &\Leftrightarrow (\mathbb{N}, (R_\beta)_{\beta \leqslant \alpha}) \vDash \varphi^*(\overline{n}) \text{ by Prop. 8.0.15} \\ &\Leftrightarrow \#\varphi^*(\overline{n}) \in R_{\alpha+1} \text{ since } \varphi^* \in \mathcal{L}_T^{\alpha+1} \end{aligned}$$

Thus $X = (T_{\alpha+1})_{\varphi^*} \in \mathcal{M}_{T_{\alpha+1}}$.

For the other direction, let $X = (T_{\alpha+1})_\varphi = \{n | \#\varphi(\overline{n}) \in T_{\alpha+1}\} \in \mathcal{M}_{R_{\alpha+1}}$, where $\varphi \in \mathcal{L}_T^{\alpha+1}$. Notice first that $n \in X \Leftrightarrow \#\varphi(\overline{n}) \in R_{\alpha+1} \Leftrightarrow (\mathbb{N}, (R_\beta)_{\beta \leqslant \alpha}) \vDash \varphi(\overline{n})$. Since by induction hypothesis $\mathcal{M}_{R_\gamma} = \mathcal{RA}_\gamma$ for all $\gamma \leqslant \beta$, Proposition 8.0.18 implies that there is a $\psi \in \mathcal{L}_2^{\alpha+1}$ such that φ defines the same set as ψ^*. Thus

$$\begin{aligned} n \in X &\Leftrightarrow \#\varphi(\overline{n}) \in R_{\alpha+1} \\ &\Leftrightarrow (\mathbb{N}, (R_\beta)_{\beta \leqslant \alpha}) \vDash \varphi(\overline{n}) \\ &\Leftrightarrow (\mathbb{N}, (R_\beta)_{\beta \leqslant \alpha}) \vDash \psi^*(\overline{n}) \\ &\Leftrightarrow (\mathbb{N}, (\mathcal{M}_{R_\beta})_{\beta \leqslant \alpha}) \vDash \psi(\overline{n}) \text{ by Prop. 8.0.15} \\ &\Leftrightarrow (\mathbb{N}, (\mathcal{RA}_\beta)_{\beta \leqslant \alpha}) \vDash \psi(\overline{n}) \text{ by I.H.} \end{aligned}$$

Hence $X \in \mathcal{RA}_{\alpha+1}$.

Limit case:

Let $X \in \mathcal{RA}_\lambda$. Thus $X \in \mathcal{RA}_\alpha$ for some $\alpha < \lambda$. By I.H. $\mathcal{RA}_\alpha = \mathcal{M}_{T_\alpha}$. Hence $X = \{n | \#\varphi(\overline{n}) \in T_\alpha\} = (T_\alpha)_\varphi$ for some $\varphi \in \mathcal{L}_T^\alpha$. By Corollary 8.0.11, $X = (T_\alpha)_\varphi = (T_\lambda)_\varphi \in \mathcal{M}_{T_\lambda}$.

For the other direction, let $X \in \mathcal{M}_{T_\lambda}$. So $X = (T_\lambda)_\varphi$ for some $\varphi \in \mathcal{L}_T^\alpha$, where $\alpha < \lambda$. By Corollary 8.0.11, $X = (T_\alpha)_\varphi \in \mathcal{M}_{T_\alpha} = \mathcal{RA}_\alpha \subseteq \mathcal{RA}_\lambda$. □

As an immediate consequence we get:

Theorem 8.0.20. *The sets definable in $L_T^{\omega_1^{CK}}$ over the standard model $(\mathbb{N}, (R_\alpha)_{\alpha \leqslant \omega_1^{CK}})$ are exactly the hyperarithmetical sets.*

9. Stratified truth

Many philosophers have noted that there are certain similarities (but also differences) between the semantic and the class paradoxes. In this chapter we have a look at some solutions to the class paradoxes, in the hope that they might help us with blocking the semantic paradoxes. In partcular, I think that Russell's work might help us a bit. Russell had a *logical* notion of class: 'a class may be defined as all the terms satisfying some propositional function.' Examples of propositional functions include 'x is a prime number' or 'x is human'. Since we can write down a formula characterizing the ordinal numbers, it seems that the ordinal numbers do form a logical collection. The logical notion of class leads itself very easily to the naive comprehension axiom scheme that causes paradox. Russell saw basically two possibilities to react to that situation:

The first one is to deny that every propositional function determines a class. Accordingly, the task is to restrict the comprehension axiom scheme in a plausible way. The difficulty then is to seperate the legitimate from the non-legitimate instances of comprehension. This is more or less the same problem that truth theorists are confronted with: we have to find restrictions on the legitimate instances of the T-schema.

The alternative is to dispense with classes altogether. This leads in the direction of the no-classes theory that we considered in section 1.2

Both routes are explored in Russell's [79] paper *On Some Difficulties in the Theory of Transfinite Numbers and Order Types* (1905). Concerning the first option, Russell suggests two ways to adress the problem of dividing the bad from the good propositional functions. The first approach is the so-called *zigzag theory*. The second is *the theory of limitation of size*. Roughly, the zigzag theory places restrictions on the complexity of propositional functions, while the limitation of size approach bans classes that are in some sense too big. Since the limitation of size approach does not seem to be applicable to the semantic paradoxes, we will only consider the zigzag approach here.

9.1. Zigzag theories

The general idea behind the zigzag approach is that only syntactially 'simple' formulae determine an extension. It is natural to think of the set of simple formulae as being closed under negation. Thus, if a class u exists, then so does its complement $\overline{u}$. Now if ψ is a condition that does not determine a class and u is an arbitrary class, then there must be members of u that do not satisfy ψ or there are members of $\overline{u}$ that do satisfy ψ. (For otherwise there would be an u such that $x \in u \leftrightarrow \psi(x)$, and hence ψ would determine a class, contrary to our assumption.) This is the zigzag property which gives the theory its name. Since there are certain seemingly simple conditions (such as $x = x$) that are satisfied by all terms, zigzag theories do not blame the size of certain classes for the paradoxes: the existence of the universal class is provable in such theories.

Russell tried to give an axiomatization of the predicate 'x is a simple propositional function' but was never quite satisfied with his results. We therefore have a look at Quine's [73] NF and then at Esser's [17] positive set theory. These might be called zigzag theories insofar as both settle on instances of the comprehension scheme that satisfy simple syntactic constraints.[1]

Russell's substitutional theory of classes (cf. section 1.2) has led him to the discovery of the theory of logical types. However, set theorists have rejected typing as an approach to the class paradoxes very much like truth theorists have rejected typing as a solution to the semantic paradoxes. Quine's idea is to drop all indices from the language, thus working with an untyped language, but to allow comprehension only for formulae that can be viewed as a 'translation' of a typed formula. More precisely, we call a formula φ *stratified* iff there is a function f that assigns natural numbers to the variables in φ in such a way that whenever $x = y$ is a subformula of φ, then $f(x) = f(y)$, and whenever $x \in y$ is a subformula of φ then $f(y) = f(x) + 1$. NF contains as axioms (besides an axiom of extensionality) all instances of the comprehension scheme

$$\exists y \forall x (x \in y \leftrightarrow \varphi(x))$$

where φ is stratified (and y does not occur free in φ).

In positive set theory [17], comprehension is allowed for positive (and generalized positive) formulae. Here, a formula is called *positive* if it belongs to the smallest class containing $\bot, x = y$ and $x \in y$ and is closed under disjunction, conjunction, and universal and existential quantification. Generalized positive formula are obtained by allowing bounded quantification. (The theory contains, in addition, a closure axiom and an axiom of infinity.)

Positive set theory and NF are both mathematically rich theories; given the strong connections between set theory and truth theory, one might flirt with the idea of zigzag theories of truth. Indeed, the theory PUTB that we have introduced in chapter

[1] However, in positive set theory the admissible instances are not closed under negation (something which is constitutive of zigzag theories as Russell understands it); hence the label zigzag theory applies here only in a derivative sense.

7 can be seen as a successful example of a zigzag theory of truth. This motivates us to investigate systems of stratified truth.

One difference between Russell's type theory of classes and Tarski's type theory of truth is that the former attaches indices not to the membership symbol but to the variables, whereas the latter attaches indices to the truth predicates but not to the variables. Dropping the indices from the Tarskian hierarchy does not work; the resulting system would simply be inconsistent. What we are looking for is something like the following.

Let $\mathcal{L}$ be some ordinary first-order language and $\mathcal{S}$ some theory in that language. Let $\mathcal{L}_{Sat}$ be the extension of $\mathcal{L}$ that is obtained by adding a binary satisfaction predicate Sat and a denumerably infinite set of variables v_i^n, one for each $i, n \in \omega$, where the superscript indicates the type or level of the variable. (We identify the variables of type 0 with the variables of the base language $\mathcal{L}$.) Let us say that $Sat(v^n, v^k)$ is well-formed iff $k = n+1$. So $\mathcal{L}_{Sat}$ is a multi-sorted first-order language. Let us also introduce, for each formula φ with exactly one free variable, a name (individual constant) $\overline{\varphi}$. Let us define the type of $\overline{\varphi}$ as the maximum of the types of the terms occurring in φ plus 1. (So a universal quantifer $\forall x^n$ may only be instantiated to $\overline{\varphi}$ if the latter is of type n.) The theory $\mathcal{S}^+$ consists of the axioms of $\mathcal{S}$ plus all sentences of the form

$$\forall x^n (Sat(x^n, \overline{\varphi}) \leftrightarrow \varphi(x^n))$$

assuming that $\overline{\varphi}$ is of type $n+1$. Then the resulting system is a typed theory of truth (or rather, satisfaction), which basically is—as is easily seen—just a version of simple type theory. Now let $\mathcal{L}'_{Sat}$ be the single-sorted first-order language obtained from $\mathcal{L}_{Sat}$ by deleting all variables of type $n > 0$. Let $\mathcal{S}^\ddagger$ be the theory in $\mathcal{L}'_{Sat}$ consisting of the axioms of $\mathcal{S}$ plus all sentences of the form

$$\forall x (Sat(x, \overline{\varphi}) \leftrightarrow \varphi(x))$$

where φ is obtained from a formula of $\mathcal{L}_{Sat}$ by dropping the type indices. Then $\mathcal{S}^\ddagger$ is an untyped theory of *stratified* truth (or satisfaction) and we might hope that it is consistent.

Let us restrict our attention for a second to a version of $\mathcal{L}_{Sat}$ with only variables of type 0 and 1. Now, letting $\mathcal{L} = \mathcal{L}_{PA}$, it is easily seen that $\mathcal{L}_{Sat}$ is basically just the language of second-order arithmetic, $\mathcal{L}_2$, augmented with set constants. In order to arrive at our desired arithmetical theory of stratified truth, we will identify $\mathcal{L}'_{Sat}$with $\mathcal{L}_T$, by stipulating that $Sat(x, y)$ is defined as $T\underset{\cdot}{s}(x, y)$ and the names $\overline{\varphi}$ are replaced by the Gödelnumerals $\ulcorner\varphi\urcorner$, but otherwise following the procedure outlined above. This means that a theory of stratified truth is obtained by adopting the uniform T-biconditionals for translations of second-order formulae.

9.2. Some systems of stratified truth

The predicate $T\underset{.}{s}(x, y)$ can be viewed as a satisfaction predicate, 'y satisfies x'. There is a close relation between the liar and Russell's paradox. Consider the formula $\neg T\underset{.}{s}(x, x)$ and let $n = \#\neg T\underset{.}{s}(x, x)$. Then $\mathsf{PA} \vdash \underset{.}{s}(\overline{n}, \overline{n}) = \ulcorner \neg T\underset{.}{s}(\overline{n}, \overline{n}) \urcorner$, that is we have produced a term t such that $\mathsf{PA} \vdash t = \ulcorner \neg Tt \urcorner$. The uniform T-biconditional for the formula 'x does not satisfy x' gives rise to a contradiction:

$$\forall x(T\ulcorner \neg T\underset{.}{s}(\dot{x}, \dot{x})\urcorner \leftrightarrow \neg T\underset{.}{s}(x, x))$$

is inconsistent over PA.

As indicated in section 9.1, an appropriate translation of higher-order formulae into formulae of $\mathcal{L}_T$ induces a stratification of the formulae that lie in the range of the translation function. Such a stratification will rule out the liar predicate as a legitimate instance of the T-schema. We first study a simple disquotational theory of truth that is obtained by adopting T-sentences for translations of second-order formulae. Though simple, that theory has remarkable deductive power. Later on, we consider some ramifications of that theory that have even more expressive power.

In order to simplify the consistency proof, we will make some assumptions about the behavior of the substitution function when applied to numbers that are not codes of formulae. So let us assume that s is a p.r. binary function with the following properties. (i) If k is the code of an $\mathcal{L}_T$-formula φ with exactly v_i free and n an arbitrary number, then $s(k, n)$ is the code of the formula which results from φ by substituting the numeral of n for all free occurrences of the variable v_i. (ii) If k is the code of a closed $\mathcal{L}_T$-formula, then $s(k, n) = k$. (iii) If k is not the code of an $\mathcal{L}_T$-formula, then $s(k, n)$ is not the code of an $\mathcal{L}_T$-formula. (iv) If neither k_1 nor k_2 are codes of $\mathcal{L}_T$-formula, then $s(k_1, n_1) = s(k_2, n_2)$ implies $k_1 = k_2$ and $n_1 = n_2$.

Now we introduce the following translation:

Definition 9.2.1. The function $^*: \mathcal{L}_2 \to \mathcal{L}_T$ is defined as follows:
$v_i^* = v_{2i}$, $X_i^* = v_{2i+1}$
$\overline{0}^* = \overline{0}$, $f(t_1, \ldots, t_n)^* = f(t_1^*, \ldots, t_n^*)$
$(s = t)^* = (s^* = t^*)$, $(\neg\varphi)^* = \neg\varphi^*$, $(\varphi \wedge \psi)^* = \varphi^* \wedge \psi^*$
$(t \in X_i)^* = T\underset{.}{s}(v_{2i+1}, t^*)$
$(\forall v_i \varphi)^* = \forall v_{2i} \varphi^*$
$(\forall X_i \varphi)^* = \forall v_{2i+1} \varphi^*$

Notice that here we depart from the standard way of dealing with the set quantifiers, which usually get relativized in the translation, as in the previous chapters. (This is the reason why we need to make some assumptions about the behavior of s when applied to non-codes.) We are now in a position to formulate our first theory of stratified truth.

Definition 9.2.2. $\mathsf{UTB}(\mathsf{Z}_2^-)$ is the theory in $\mathcal{L}_T$ whose axioms comprise those of PAT and all instances of the following axiom scheme:

$$\forall x_1 \dots \forall x_n (T\ulcorner \varphi(\dot{x_1}, \dots, \dot{x_n})\urcorner \leftrightarrow \varphi(x_1, \dots, x_n)),$$

where φ is the translation of an $\mathcal{L}_2$-formula ψ that contains no free set variables, and $x_1, \dots, x_n$ is an exhaustive list of all the free variables in φ.

We will show in section 9.3 that $\mathsf{UTB}(\mathsf{Z}_2^-)$ is ω-consistent. However, if we allow ψ (in the above definition) to contain free set variables, then the resulting system would be inconsistent. For example, if x is an odd and y an even variable, then $\neg T\underset{\cdot}{s}(x, y)$ is the translation of a second-order formula of the form $\neg(y \in X)$. Thus

$$\forall x \forall y (T\ulcorner \neg T\underset{\cdot}{s}(\dot{x}, \dot{y})\urcorner \leftrightarrow \neg T\underset{\cdot}{s}(x, y))$$

would be an axiom of the truth theory. But the above axiom is inconsistent over PAT, because both quantifiers can be instantiated to the term $\ulcorner \neg T\underset{\cdot}{s}(x, x)\urcorner$.

In order to measure the proof-theoretic strength of $\mathsf{UTB}(\mathsf{Z}_2^-)$, we need the following definition.

Definition 9.2.3. Z_2^- is the theory in $\mathcal{L}_2$ that contains in addition to the axioms of PA all comprehension axioms

$$\forall \vec{y} \exists X \forall x (x \in X \leftrightarrow \varphi(\vec{y}, x)),$$

where $\vec{y} = y_1, \dots, y_n$ and $\varphi(\vec{y}, x)$ is a formula of $\mathcal{L}_2$ *without free set variables*, and the induction axiom

$$\overline{0} \in X \wedge \forall x (x \in X \rightarrow x + \overline{1} \in X) \rightarrow \forall x (x \in X).$$

$\mathsf{UTB}(\mathsf{Z}_2^-)$ derives the uniform T-biconditionals for (the translation of) any parameter-free second-order formula. (In particular, it derives the uniform T-biconditionals for all sentences of $\mathcal{L}_{PA}$.) But the uniform T-biconditionals imply the corresponding comprehension axioms. We therefore get:

Proposition 9.2.4. *If φ is a theorem of Z_2^-, then $\mathsf{UTB}(\mathsf{Z}_2^-) \vdash \varphi^*$.*

Proof. It suffices to show that every axiom of Z_2^- is provable in $\mathsf{UTB}(\mathsf{Z}_2^-)$. The claim is trivial if φ is a first-order axiom of PA. Now we derive the comprehension axioms of Z_2^- in $\mathsf{UTB}(\mathsf{Z}_2^-)$. Therefore, let $\varphi(\vec{z}, u) \in \mathcal{L}_2$, with exactly $\vec{z}, u$ free. Let $\varphi^*(\vec{z^*}, u^*)$ be the translation of $\varphi(\vec{z}, u)$. The following is an axiom of $\mathsf{UTB}(\mathsf{Z}_2^-)$:

$$\forall \vec{z^*} \forall u^* (T\ulcorner \varphi^*(\dot{\vec{z}}^*, \dot{u^*})\urcorner \leftrightarrow \varphi^*(\vec{z^*}, u^*)).$$

Instantiating the outmost quantifiers we get

$$\forall u^*(T\ulcorner\varphi^*(\dot{\vec{z}^*},\dot{u}^*)\urcorner \leftrightarrow \varphi^*(\vec{z^*},u^*))),$$

which is equivalent to

$$\forall u^*(T\underset{\cdot}{s}(\ulcorner\varphi(\dot{\vec{z}^*},u^*)\urcorner,u^*) \leftrightarrow \varphi^*(\vec{z^*},u^*))).$$

Existential weakening yields

$$\exists w\forall u^*(T\underset{\cdot}{s}(w,u^*) \leftrightarrow \varphi^*(\vec{z^*},u^*))),$$

where w is an odd variable. Now we re-introduce the universal quantifiers:

$$\forall\vec{z^*}\exists w\forall u^*(T\underset{\cdot}{s}(w,u^*) \leftrightarrow \varphi^*(\vec{z^*},u^*)).$$

But this is just the translation of the comprehension axiom for $\varphi(\vec{z},u)$. The induction axiom of Z_2^- translates as follows:

$$T\underset{\cdot}{s}(x,\overline{0}) \wedge \forall y(T\underset{\cdot}{s}(x,y) \to T\underset{\cdot}{s}(x,y+\overline{1})) \to \forall yT\underset{\cdot}{s}(x,y),$$

where x is an odd and y is an even variable. And this is simply an instance of induction in PAT. Notice that $\mathsf{UTB}(\mathsf{Z}_2^-)$ also proves the translation of the second-order induction *scheme*. □

Proposition 9.2.4 shows that Z_2^- is proof-theoretically reducible to $\mathsf{UTB}(\mathsf{Z}_2^-)$. So far, the strongest systems that we have seen were Cantini's system VF and my variant VFG, which have the same arithmetical consequences as the system ID_1. The proof-theoretic ordinal of ID_1 is the Bachmann-Howard ordinal, which is also the ordinal of Kripke-Platek set theory with an axiom of infinity anf that of parameter-free Π^1_1-comprehension ($\Pi^1_1-\mathsf{CA}_0^-$). And Z_2^- does not only contain $\Pi^1_1-\mathsf{CA}_0^-$, but $\Pi^1_n-\mathsf{CA}_0^-$ for every $n \in \omega$. Therefore, $\mathsf{UTB}(\mathsf{Z}_2^-)$ exceeds all of the truth theories in deductive power *by far*.

Even though most (model-theoretic) arguments in formal philosophy are carried out with Zermelo-Fraenkel set theory ZF as the background theory, most of these arguments can actually be carried out in comparatively weak subsystems of Z_2. This applies in particular to the model constructions of Kripke [53], Herzberger [42] or Field [24]. (For example, Field's construction can be carried out within $\Pi^1_3 - \mathsf{CA}$. Cf. Welch [92].)

Thus, the main appeal of strong theories of truth like $\mathsf{UTB}(\mathsf{Z}_2^-)$ (and extensions of it that we will consider in a moment) stems from the fact that they allow us to really engage in semantics—*within* the object language. By this I mean that we are able to formalize important semantic concepts from the literature on truth within the object language $\mathcal{L}_T$ and prove relevant facts about them. We illustrate this with a very simple example:

Proposition 9.2.5. $\mathsf{UTB}(\mathsf{Z}_2^-)$ *proves the existence of the minimal Strong Kleene fixed-point.*

Proof. There is a second-order formula $\zeta(x, Y)$ with exactly x and Y free (and without any bound set variables) such that Y is a Kripke fixed point iff $\forall x(x \in Y \leftrightarrow \zeta(x, Y))$. (For details, see Halbach [38], pp. 203-204.) Thus the formula

$$\forall Y(\forall x(x \in Y \leftrightarrow \zeta(x, Y)) \to \zeta(u, Y))$$

expresses that u is a member of the minimal Kripke fixed point. (Cf. Moschovakis [62] for details on fixed points of positive inductive definitions.) The displayed formula is a parameter-free Π_1^1-formula. Thus Z_2^- proves the existence of the minimal Kripke fixed point, i.e. it proves

$$\exists X \forall u(u \in X \leftrightarrow \forall Y(\forall x(x \in Y \leftrightarrow \zeta(x, Y) \to \zeta(u, Y))),$$

and by Proposition 9.2.4, $\mathsf{UTB}(\mathsf{Z}_2^-)$ proves the translation of the last formula. □

Thus, we can formalize within $\mathsf{UTB}(\mathsf{Z}_2^-)$ that x is grounded (in the sense of Kripke), because, by definition, x is grounded iff x is in the minimal Kripke fixed-point.

Despite its enormous expressive power, $\mathsf{UTB}(\mathsf{Z}_2^-)$ fails to prove certain sentences that we expect a good truth theory to prove. For example, it does not prove the T-biconditional for the simple sentence $T\ulcorner 1 = 1\urcorner$ (because the latter is not a translation of a second-order formula). We can improve our theory as follows.

Definition 9.2.6. Let A be the smallest set X such that (i) whenever φ is an $\mathcal{L}_2$-formula without free set variables, then $\#\varphi^* \in X$, (ii) whenever φ is a theorem of PAT, then $\#\varphi \in X$, and (iii) whenever $\#\varphi, \#\psi \in X$, then $\#T\ulcorner\varphi\urcorner, \#\neg\varphi, \#(\varphi \wedge \psi) \in X$. Let $\mathsf{UTB}(\mathsf{Z}_2^-)^+$ be PAT plus all instances of the *relativized* uniform T-schema:

$$\psi(\ulcorner\varphi(\vec{x})\urcorner) \to \forall\vec{x}(T\ulcorner\varphi(\dot{\vec{x}})\urcorner \leftrightarrow \varphi(\vec{x})),$$

where $\psi(x)$ is an $\mathcal{L}_{PA}$-formula that defines A, φ is an $\mathcal{L}_T$-formula and $\vec{x} = x_1, \ldots, x_n$ is an exhaustive list of all the free variables in φ.

Clearly, $\mathsf{UTB}(\mathsf{Z}_2^-)$ is a subtheory of $\mathsf{UTB}(\mathsf{Z}_2^-)^+$ and the latter overcomes the problem of the former. Since $1 = 1$ is an element of A, $T\ulcorner 1 = 1\urcorner$ will also be an element of A, hence $T\ulcorner T\ulcorner 1 = 1\urcorner\urcorner \leftrightarrow T\ulcorner 1 = 1\urcorner$ will be an axiom of $\mathsf{UTB}(\mathsf{Z}_2^-)^+$. Of course, we can also prove the T-sentence for sentences that result from iterated applications of the truth predicate to $1 = 1$ (or any other arithmetical sentence).

We have critizised Horwich's notion of grounding (cf. section 6.4) because it does not render $\forall x(Sent_{PA}(x) \to Tx \vee T\dot{\neg}x)$ as a legitimiate instance of the T-schema. $\mathsf{UTB}(\mathsf{Z}_2^-)^+$ shares that deficiency. There are at least two ways in which we can further improve our theories. First, since we can formalize second-order notions within our object language, we might relativize the T-schema to sets that are no longer arithmetically definable but have higher complexity. For example, we have seen (Proposition 9.2.5) that the notion 'x is grounded' (in the Strong Kleene sense) can be formalized within a theory that contains the T-biconditionals for translations of second-order formulae. Thus we might add to such a theory all instances of the relativized uniform T-schema

$$\ulcorner\varphi(\vec{x})\urcorner \text{ is grounded} \to \forall\vec{x}(T\ulcorner\varphi(\dot{\vec{x}})\urcorner \leftrightarrow \varphi(\vec{x})).$$

Although that seems to be an interesting option, I won't explore it any further here. Another option would be to add a primitive predicate $Acc(x)$, intended to express 'x is an acceptable instance of the T-schema', and to give a simultaneous axiomatization of acceptability and truth, similar to our axiomatizations of grounded truth (chapter 6). In what follows, we give an example of such a theory. We need the following definitions.

Let Rel represent the binary relation that holds between (the code of) a closed $\mathcal{L}_T$-formula φ and (the code of) a sequence of $\mathcal{L}_{PA}$-formulae $(\psi_1(x), \ldots, \psi_n(x))$ iff every subformula of φ of the form Tt occurs in the context $\psi_i(t) \wedge Tt$ within φ for some $0 < i \leqslant n$. Let $seq(x)$ express that x is (the code of) a sequence. We write $\forall\sigma$ instead of $\forall x(seq(x) \to \ldots)$. We write $lh(\sigma)$ for the length of σ and $\sigma(u)$ for the u-th member of σ. Let $Fm_T(x)$ represent the set (of codes) of $\mathcal{L}_T$-formulae and let $Prov_{PAT}(x)$ be a standard provability predicate for PAT. Finally, let $Trsl(x)$ represent the set of (codes of) translations of second-order formulae without free set variables.

Definition 9.2.7. We adopt the following *axioms of acceptability*:

1. $\forall x(Fm_T(x) \to (Prov_{PAT}(x) \to Acc(x)))$,
2. $\forall x(Fm_T(x) \to (Trsl(x) \to Acc(x)))$,
3. $\forall x(Fm_T(x) \to (Acc(x) \to Acc(\underset{\cdot}{T}x)))$,
4. $\forall x(Fm_T(x) \to (Acc(x) \to Acc(\underset{\cdot}{\neg}x)))$,
5. $\forall x\forall y(Fm_T(x\underset{\cdot}{\wedge}y) \to (Acc(x) \wedge Acc(y) \to Acc(x\underset{\cdot}{\wedge}y)))$,
6. $\forall x\forall\sigma(Rel(x,\sigma) \wedge \forall u < lh(\sigma)\forall z(T\underset{\cdot}{s}(\sigma(u), z) \to Acc(z)) \to Acc(x))$.

Definition 9.2.8. $\mathsf{UTB}(\mathsf{Z}_2^-)^\ddagger$ is the theory in $\mathcal{L}_T^{Acc}$ whose axioms comprise those of PAT (with induction expanded to the full language), Axioms 1-6 of Definition 9.2.7, and all instances of the following axiom scheme:

$$Acc(\ulcorner\varphi(\vec{x})\urcorner) \to \forall\vec{x}(T\ulcorner\varphi(\dot{\vec{x}})\urcorner) \leftrightarrow \varphi(\vec{x})),$$

where $\varphi(\vec{x})$ is an $\mathcal{L}_T$-formula with exactly $\vec{x} = x_1, \ldots, x_n$ free.

$\mathsf{UTB}(\mathsf{Z}_2^-)^\ddagger$ proves the T-biconditional for $\forall x(Sent_{PA}(x) \to Tx \vee T\underset{\cdot}{\neg}x)$. More generally, let us show that all sentences of the Tarskian hierarchy of truth (i.e. their translations) are acceptable. The formulae of RT can be translated into $\mathcal{L}_T$ in a straightforward way.

Definition 9.2.9. The sublanguages L_α of $\mathcal{L}_T$ are defined by recursion over the ordinals up to ϵ_0. L_0 is just $\mathcal{L}_{PA}$. For $0 < \alpha < \epsilon_0$, φ is a formula of the language L_α iff there are $\beta_1, \ldots \beta_n < \alpha$ such that every occurrence of a subformula Tt of φ occurs in the context $Sent(\overline{\beta_i}, t) \wedge Tt$ for some $0 < i \leqslant n$, where $Sent(\overline{\beta_i}, x)$ represents the set of L_{β_i}-sentences.

We can define 'x is a sentence of L_α' as follows, using Kleene's recursion theorem (where $OT(x)$ means that x is an ordinal term):

$$Sent(\alpha, x) \leftrightarrow [OT(\alpha) \wedge \exists \sigma, \tau < x(lh(\sigma) = lh(\tau) \wedge Rel(x, \sigma) \wedge$$

$$\wedge \forall u < lh(\tau)(OT(\tau(u)) \wedge \tau(u) \prec \alpha \wedge \sigma(u) = \ulcorner Sent(\tau(\dot{u}), v_0)\urcorner))]$$

Using transfinite induction, one can then show:

Proposition 9.2.10. *For all $\delta < \epsilon_0$,*

$$\mathsf{UTB}(\mathsf{Z}_2^-)^\ddagger \vdash \forall \zeta \prec \overline{\delta} \forall x(Sent(\zeta, x) \to Acc(x)).$$

Proof. Let $\varphi(v)$ be the formula $\forall x(Sent(v, x) \to Acc(x))$. PAT proves transfinite induction for every $\delta \prec \epsilon_0$, i.e. for all $\delta \prec \epsilon_0$ PAT proves:

$$\forall \alpha(\forall \beta \prec \alpha \varphi(\beta) \to \varphi(\alpha)) \to \forall \zeta \prec \overline{\delta} \varphi(\zeta).$$

So assume

$$\forall \beta \prec \alpha \forall x(Sent(\beta, x) \to Acc(x)). \quad (I.H.)$$

Then it suffices to show that

$$\forall x(Sent(\alpha, x) \to Acc(x)).$$

Therefore let x be given and assume $Sent(\alpha, x)$. Then PA proves

$$OT(\alpha) \wedge \exists \sigma, \tau < x(lh(\sigma) = lh(\tau) \wedge Rel(x, \sigma) \wedge$$

$$\wedge \forall u < lh(\tau)(OT(\tau(u)) \wedge \tau(u) \prec \alpha \wedge \sigma(u) = \ulcorner Sent(\tau(\dot{u}), v_0)\urcorner)).$$

Let $\sigma, \tau < x$ and $u < lh(\tau) = lh(\sigma)$ be as above. Because the formula $Sent(\tau(u), v_0)$ is arithmetical and $\mathsf{UTB}(\mathsf{Z}_2)^\ddagger$ proves the uniform T-biconditionals for all $\mathcal{L}_{PA}$-formulae, we get

$$\forall u \forall v_0(T\ulcorner Sent(\tau(\dot{u}), \dot{v_0})\urcorner \leftrightarrow Sent(\tau(u), v_0)). \tag{9.1}$$

Since $\tau(u) \prec \alpha$, (I.H.) yields

$$\forall z(Sent(\tau(u), z) \rightarrow Acc(z)). \tag{9.2}$$

Because $\sigma(u) = \ulcorner Sent(\tau(\dot{u}), v_0)\urcorner$, (9.1) and (9.2) yield

$$\forall z(T\underset{\cdot}{s}(\sigma(u), z) \rightarrow Acc(z)).$$

Since this holds for all $u < lh(\sigma)$, Axiom 6 of Definition 9.2.7 yields $Acc(x)$. □

The ramifications introduced so far add more truth-theoretic content but, most likely, these systems do not prove more arithmetical sentences than our first theory, $\mathsf{UTB}(\mathsf{Z}_2^-)$. A theory with higher proof-theoretic strength might be obtained by adopting T-biconditionals for (translations of) sentences of higher-order arithmetic.

Definition 9.2.11. 1. $\mathcal{L}_\omega$ is the language of Peano arithmetic augmented with a binary relation symbol $\in$ plus countably many *indexed* set variables $X_1^n, X_2^n, X_3^n, \ldots$, for every index $n \in \omega \setminus \{0\}$. This gives us new formulae of the form $t \in X^1, X^n \in X^{n+1}$, and $\forall X^n \varphi$. $\mathcal{L}_\omega$ is a many-sorted first-order language with usual quantifier rules.

2. The language $\mathcal{L}_\omega$ can be regarded as a sublanguage of $\mathcal{L}_T$ by the following stipulation. Let (a, b) be the code of the ordered pair a, b under the Cantor pairing function. Let $\varphi \in \mathcal{L}_\omega$. Replace any occurrence of v_i by $v_{(0,i)}$, every occurrence of X_k^n by $v_{(n,k)}$, and every occurrence of $\in$ by $T\underset{\cdot}{s}(\cdot,\cdot)$. We denote the result of this replacement by φ^{**}.

3. Z_ω^- consists of the axioms of PA with induction extended to $\mathcal{L}_\omega$ plus all formulae of the form
$$\forall \vec{Z} \exists X^{n+1} \forall X^n (X^n \in X^{n+1} \leftrightarrow \varphi(X^n, \vec{Z})),$$
where φ is an $\mathcal{L}_\omega$-formula which does not contain X^{n+1} free and in which no free variable occurs on the right-hand side of the symbol $\in$. (For $n = 0$ we assume that X^n is some number variable v_i).

The system Z_ω is obtained from Z_ω^- by allowing free set parameters on the right-hand side of the symbol $\in$; this system is roughly equivalent to the simple theory of types with an axiom of infinity. The latter is basically a simplified version of Russell and Whitehead's *Principia Mathematica.*

Definition 9.2.12. $\mathsf{UTB}(\mathsf{Z}_\omega^-)$ is the theory in $\mathcal{L}_T$ whose axioms are those of PAT plus all sentences of the form

$$\forall x_1 \ldots \forall x_n (T\ulcorner\varphi(\dot{x_1}, \ldots, \dot{x_n})\urcorner \leftrightarrow \varphi(x_1, \ldots, x_n)),$$

where φ is the translation of an $\mathcal{L}_\omega$-formula ψ, $x_1, \ldots, x_n$ is an exhaustive list of all the free variables in φ, and no free variable in ψ occurs on the right-hand side of the symbol $\in$.

Allowing ψ (in the above definition) to contain free variables on the right-hand side of the symbol $\in$ renders the system inconsistent. For in that event, the formula

$$\neg T\underset{.}{s}(x, \ulcorner\neg T\underset{.}{s}(x,x)\urcorner)$$

(which is a translation of the second-order formula $\ulcorner\neg T\underset{.}{s}(x,x)\urcorner \in X^1$) would be a legitimate instance of the T-schema.

Proposition 9.2.13. *If φ is a theorem of Z_ω^-, then $\mathsf{UTB}(\mathsf{Z}_\omega^-) \vdash \varphi^{**}$.*

Proof. Similar to the proof of Proposition 9.2.4. □

9.3. Consistency

The goal of this section is to prove the existence of standard models for the theories $\mathsf{UTB}(\mathsf{Z}_2^-)$ and $\mathsf{UTB}(\mathsf{Z}_2^-)^\ddagger$ (and therefore for $\mathsf{UTB}(\mathsf{Z}_2^-)^+$ too).[2] In order to do so, we again associate second-order structures with truth-sets. We need to generalize our old definitions.

Definition 9.3.1. Let $S \subseteq \omega$.

1. $S_k = \{n | s(k,n) \in S\}$
2. $\mathcal{M}_S = \{S_k | k \in \omega\}$.

The main difference between this and our old definition (7.1.2) is that the sets S_k are now also defined for k that do not code a formula. (The reason why we need to do this lies in the fact that this time we did not relativize the set quantifiers in the translation of $\mathcal{L}_2$ into $\mathcal{L}_T$.) Note that if $k = \#\varphi$ for some formula φ, then our above definition of S_φ coincides with our earlier definition. Again, we obtain a Translation Lemma.

If h is a variable assignment for $(\mathbb{N}, \mathcal{M}_S)$, define the assignment h^* for $(\mathbb{N}, S)$ by $h^*(v_{2i}) = h(v_i)$ and $h^*(v_{2i+1}) = \min\{k \mid S_k = h(X_i)\}$.

Proposition 9.3.2. *Let h be an assignment for $(\mathbb{N}, \mathcal{M}_S)$. Then $t^{(\mathbb{N},\mathcal{M}_S),h} = t^{*(\mathbb{N},S),h^*}$ for all number terms t of $\mathcal{L}_2$.*

Proposition 9.3.3. *(Translation Lemma) Let $S \subseteq \omega$, let $\varphi(x, \vec{y}, \vec{X}) \in \mathcal{L}_2$, and let h be an assignment for $(\mathbb{N}, \mathcal{M}_S)$. Then:*

$$(\mathbb{N}, \mathcal{M}_S), h \vDash \varphi \Leftrightarrow (\mathbb{N}, S), h^* \vDash \varphi^*$$

[2] I do not know whether $\mathsf{UTB}(\mathsf{Z}_\omega^-)$ is consistent, although I assume so.

Proof. By induction on the complexity of formulae. The case $s = t$ is trivial.

Consider $t \in X_i$, where t is any term. Let $t^{(\mathbb{N},\mathcal{M}_S),h} = n$ and $h(X_i) = A$. There's a k such that $k = \min\{m \mid S_m = A\}$. Then

$$\begin{aligned}(\mathbb{N}, \mathcal{M}_S), h \vDash t \in X_i &\Leftrightarrow n \in A \\ &\Leftrightarrow n \in S_k \\ &\Leftrightarrow s(k, n) \in S \\ &\Leftrightarrow (\mathbb{N}, S) \vDash T\underset{\cdot}{s}(\overline{k}, \overline{n}) \\ &\Leftrightarrow (\mathbb{N}, S), h^* \vDash T\underset{\cdot}{s}(v_{2i+1}, t^*)\end{aligned}$$

The cases $\neg\psi, \psi \wedge \chi$ and $\forall x\psi$ follow easily from the I.H. Finally, consider $\forall X_i\psi$ and let $M = \{k|\forall m(S_m = S_k \to k \leqslant m)\}$.

$$\begin{aligned}(\mathbb{N}, \mathcal{M}_S), h \vDash \forall X_i\psi &\Leftrightarrow \forall A \in \mathcal{M}_S : (\mathbb{N}, \mathcal{M}_S), h(A : X_i) \vDash \psi && (9.3)\\ &\Leftrightarrow \forall k \in \omega : (\mathbb{N}, \mathcal{M}_S), h(S_k : X_i) \vDash \psi && (9.4)\\ &\Leftrightarrow \forall k \in M : (\mathbb{N}, \mathcal{M}_S), h(S_k : X_i) \vDash \psi && (9.5)\\ &\Leftrightarrow \forall k \in M : (\mathbb{N}, S), h^*(k : v_{2i+1}) \vDash \psi^* && (9.6)\\ &\Leftrightarrow \forall k \in \omega : (\mathbb{N}, S), h^*(k : v_{2i+1}) \vDash \psi^* && (9.7)\\ &\Leftrightarrow (\mathbb{N}, S), h^* \vDash \forall v_{2i+1}\psi^* && (9.8)\end{aligned}$$

The implication from (9.5) to (9.4) follows from the definition of M and the extensionality of sets. The equivalence between (9.5) and (9.6) is given by the inductive hypothesis, since $[h(S_k : X_i)]^* = h^*(k : v_{2i+1})$ for every minimal k. The step from (9.6) to (9.7) is justified because in translated formulae, such as ψ^*, v_{2i+1} occurs only in contexts of the form $T\underset{\cdot}{s}(v_{2i+1}, t)$. By the definition of the sets S_k, if $S_m = S_k$ then

$$(\mathbb{N}, S), h' \vDash T\underset{\cdot}{s}(\overline{k}, t) \leftrightarrow T\underset{\cdot}{s}(\overline{m}, t),$$

for any term t and assignment h'. □

Now we may prove:

Proposition 9.3.4. $\mathsf{UTB}(\mathsf{Z}_2^-)$ *has an ω-model.*

Proof. Let $\mathcal{RA}$ be the collection of *ramified analytic sets* (up to level β_0, which is countable). It is well-known that $(\mathbb{N}, \mathcal{RA})$ is a countable $\mathcal{L}_2$-structure that is closed under second-order definability with parameters from $\mathcal{RA}$, hence a model of Z_2 and therefore of Z_2^- (cf. [69]). Let $enum : (\omega \setminus Fm_{\mathcal{L}_T}) \to \mathcal{RA}$ be bijective.

Let $s(k, n) \in S$ iff

1. k does not code an $\mathcal{L}_T$-formula and $n \in enum(k)$; or

2. $\varphi(x)$ is an $\mathcal{L}_2$-formula with exactly x free, k the code of $\varphi^*(x^*)$, and $(\mathbb{N}, \mathcal{RA}) \vDash \varphi(\overline{n})$.

Then it is easily seen that $\mathcal{M}_S = \mathcal{RA}$. We now show that $(\mathbb{N}, S)$ validates the theory $\mathsf{UTB}(\mathsf{Z}_2^-)$.

Let φ be an $\mathcal{L}_2$-formula without free set variables. Assume that $\vec{y}, x$ is an exhaustive list of all free variables in φ. Let $\vec{m}$ be given. We have to show that

$$(\mathbb{N}, S) \vDash \forall x^*(T\underset{\cdot}{s}(\ulcorner\varphi^*(\vec{\overline{m}}, x^*)\urcorner, x^*) \leftrightarrow \varphi^*(\vec{\overline{m}}, x^*)).$$

Let n be given and assume $(\mathbb{N}, S) \vDash \varphi^*(\vec{\overline{m}}, \overline{n})$. By Proposition 9.4.6, $(\mathbb{N}, \mathcal{M}_S) \vDash \varphi(\vec{\overline{m}}, \overline{n})$. Since $\mathcal{M}_S = \mathcal{RA}$, we have $s(\#\varphi^*(\vec{\overline{m}}, x^*), n) \in S$ by (ii). Thus $(\mathbb{N}, S) \vDash T\underset{\cdot}{s}(\ulcorner\varphi^*(\vec{\overline{m}}, x^*)\urcorner, \overline{n})$. The argument for the other direction of the claim is similar. □

We close this section by outlining a consistency proof for one of the more expressive systems.

Proposition 9.3.5. $\mathsf{UTB}(\mathsf{Z}_2^-)^\ddagger$ *has an* ω*-model.*

Proof. (Sketch) The predicate Acc will be interpreted by the fixed point of an inclusive hierarchy that is defined by transfinite induction as follows. Let Θ_0 be the set consisting exactly of (the codes of) all theorems of PAT and all translations of second-order formulae that do not contain free set variables. Let $\Theta_{\alpha+1}$ be the smallest superset X of Θ_α such that (i) whenever $\#\varphi, \#\psi \in \Theta_\alpha$ then $\#\neg\varphi, \#(\varphi \wedge \psi)$ and $\#T\ulcorner\varphi\urcorner \in X$, and (ii) whenever $\psi_1, \ldots, \psi_n$ are $\mathcal{L}_{PA}$-formulae with $\psi_i^{\mathbb{N}} \subseteq \Theta_\alpha$ (for all $i \leqslant n$) and φ is relativized to $(\psi_1, \ldots, \psi_n)$, then $\#\varphi \in X$. At limit points we take unions. Let Θ denote the fixed point of that hierarchy. Every code $\#\varphi$ in Θ will be assigned an ordinal rank denoting the least level at which $\#\varphi$ enters the hierarchy.

Let $S^0 = S \cup \{\#\varphi | \varphi \in \mathcal{L}_T, \mathsf{PAT} \vdash \varphi\}$, where S is defined as in the proof of Proposition 9.3.4. It is not hard to prove that $(\mathbb{N}, S^0)$ validates the T-biconditionals for all sentences of rank 0 (i.e. translations of second-order formulae without free set variables plus all theorems of PAT). Let $S^{\alpha+1}$ be the set of all sentences $\#\varphi \in \Theta_{\alpha+1}$ that are true in $(\mathbb{N}, S^\alpha)$. At limit points we take unions. The fixed point of the S_α-hierarchy will validate the T-biconditionals for all acceptable sentences (and since the model is standard, it will therefore validate the uniform T-biconditionals for all acceptable formulae). For one can show (for each α) that all sentences $\#\varphi \in \Theta_\alpha$ have the same truth value in all $(\mathbb{N}, S^\beta)$ for all $\beta \geqslant \alpha$. In order to show this, one may verify that for all $\alpha, \mathcal{M}_{S_\alpha} = \mathcal{RA}$. This guarantees that all sentences of rank 0 have the same truth value in all models $(\mathbb{N}, S_\alpha)$. Sentences of rank > 0 preserve their truth value because they depend (in the sense of Leitgeb [55]) on the set of sentences of lower rank. □

9.4. Comprehension with parameters

Can we design theories of truth that are able to derive comprehension axioms for formulae that contain free set variables? There is indeed such a method, although I don't think that the resulting systems are very attractive. For the sake of completeness, we sketch the method. For simplicity, we focus on the question of how to interpret ACA (see the appendix for a definition) in a disquotational theory.[3] The method can be generalized to yield an interpretation of Z_2.

A natural thought is to relativize the higher-order quantifiers in the translation, in order to exclude the liar predicate from the range of the quantifiers. First, let us show that the predicate $Fm^1_T(x)$—'x is a formula of $\mathcal{L}_T$ with exactly one free variable'—won't do the job.

Proposition 9.4.1. *The following schema is inconsistent with* PA.

$$\forall\vec{z}\forall\vec{y}(Fm^1_T(\vec{y})\rightarrow\forall x(T\ulcorner\varphi^*(\dot{x},\dot{\vec{z}},\dot{\vec{y}})\urcorner\leftrightarrow\varphi^*(x,\vec{z},\vec{y}))) \tag{9.9}$$

where $\varphi(x,\vec{z},\vec{Y})\in\mathcal{L}_2$ is arithmetical and φ^* is its translation, where now set quantifiers get relativized to the predicate $Fm^1_T(x)$.

Proof. Consider the $\mathcal{L}_2$-formula $\neg(x\in Y)$. This is arithmetical, and its translation is $\neg T\underset{\cdot}{s}(y,x)$. Let $\overline{n}$ be $\ulcorner\neg T\underset{\cdot}{s}(z,z)\urcorner$. Notice that PA proves $Fm^1_T(\overline{n})$. Applying (9.9) to $\neg T\underset{\cdot}{s}(y,x)$ and unpacking notation we get

$$\forall y(Fm^1_T(y)\rightarrow\forall x(T\underset{\cdot}{s}(\underset{\cdot}{s}(\ulcorner\neg T\underset{\cdot}{s}(y,x)\urcorner,y),x)\leftrightarrow\neg T\underset{\cdot}{s}(y,x)))$$

Then:

$$\begin{aligned}
T\underset{\cdot}{s}(\underset{\cdot}{s}(\ulcorner\neg T\underset{\cdot}{s}(y,x)\urcorner,y),x)&\leftrightarrow\neg T\underset{\cdot}{s}(y,x)\\
T\underset{\cdot}{s}(\underset{\cdot}{s}(\ulcorner\neg T\underset{\cdot}{s}(y,x)\urcorner,\overline{n}),x)&\leftrightarrow\neg T\underset{\cdot}{s}(\overline{n},x)\\
T\underset{\cdot}{s}(\ulcorner\neg T\underset{\cdot}{s}(\overline{n},x)\urcorner,x)&\leftrightarrow\neg T\underset{\cdot}{s}(\overline{n},x)\\
T\underset{\cdot}{s}(\ulcorner\neg T\underset{\cdot}{s}(\overline{n},x)\urcorner,\overline{n})&\leftrightarrow\neg T\underset{\cdot}{s}(\overline{n},\overline{n})\\
T\ulcorner\neg T\underset{\cdot}{s}(\overline{n},\overline{n})\urcorner&\leftrightarrow\neg T\underset{\cdot}{s}(\overline{n},\overline{n})\\
T\ulcorner\neg T\underset{\cdot}{s}(\overline{n},\overline{n})\urcorner&\leftrightarrow\neg T\ulcorner\neg T\underset{\cdot}{s}(\overline{n},\overline{n})\urcorner
\end{aligned}$$

□

Let $\mathcal{L}_2^+$ be the minimal extension of $\mathcal{L}_2$ that is closed under the following rule: if $\varphi\in\mathcal{L}_2^+$ is a formula with exactly x_0 free, then $\mathcal{L}_2^+$ contains a set constant $\overline{S_\varphi}$ (with $t\in\overline{S_\varphi}$ being an atomic formula of $\mathcal{L}_2^+$). More precisely, we define $L^0_2=\mathcal{L}_2$, $L^{n+1}_2=L^n_2\cup\{\overline{S_\varphi}|\varphi\in L^n_2\}$. Then we let $\mathcal{L}_2^+=\bigcup_{n\in\omega}L_n$.

[3] I thank Lavinia Picollo for her help in clarifying some of the issues in this section.

Proposition 9.4.2. *The recursion theorem for primitive recursive functions yields the existence of a primitive recursive translation function $\tau : \mathcal{L}_2^+ \to \mathcal{L}_T$ such that:*

$$\tau(\varphi) = \begin{cases} x_{2i}, & \textit{if } t \textit{ is } x_i \\ x_{2i+1}, & \textit{if } t \textit{ is } X_i \\ f(\tau(t_1), \dots, \tau(t_n)), & \textit{if } t \textit{ is } f(t_1, \dots, t_n) \\ \overline{\#\tau(\varphi)}, & \textit{if } t \textit{ is } \overline{S_\varphi} \\ \tau(s) = \tau(t), & \textit{if } \varphi := s = t \\ T\underset{.}{s}(x_{2i+1}, \tau(t)) & \textit{if } \varphi := t \in X_i \\ T\underset{.}{s}(\ulcorner\tau(\varphi)\urcorner, \tau(t)) & \textit{if } \varphi := t \in \overline{S_\varphi} \\ \neg\tau(\psi) & \textit{if } \varphi := \neg\psi \\ \tau(\psi) \wedge \tau(\chi) & \textit{if } \varphi := \psi \wedge \chi \\ \forall x_{2i}\tau(\psi) & \textit{if } \varphi := \forall x_i\psi \\ \forall x_{2i+1}(\exists y(x_{2i+1} = \underset{.}{\tau}(y) \wedge Fm_T^1(x_{2i+1})) \to \tau(\psi)) & \textit{if } \varphi := \forall X_i\psi \end{cases}$$

where $\underset{.}{\tau}$ is a function symbol for τ in $\mathcal{L}_T$.

The above function is well-defined. For τ is well-defined on $L_2^0 = \mathcal{L}_2$, and assuming that τ is well-defined on L_2^n, it is easy to show that τ is well-defined on L_2^{n+1}, too.

We abbreviate $\exists y(x = \underset{.}{\tau}(y) \wedge Fm_T^1(x))$ by $Trsl(x)$.

Definition 9.4.3. The theory $\mathsf{UTB(ACA)}$ is given by the axioms of PAT plus all instances of the following scheme:

$$\forall\vec{y}\forall\vec{z}(Trsl(\vec{y}) \to \forall x(T\underset{.}{s}(\ulcorner\tau(\varphi)(x, \dot{\vec{z}}, \dot{\vec{y}})\urcorner, x) \leftrightarrow \tau(\varphi)(x, \vec{z}, \vec{y}))) \tag{9.10}$$

where $\varphi(x, \vec{z}, \vec{Y}) \in \mathcal{L}_2$ is arithmetical.

Proposition 9.4.4. ACA *is relatively interpretable in* $\mathsf{UTB(ACA)}$.

Proof. Let $\varphi \in \mathcal{L}_2$. We have to show that the translation of the comprehension axiom for φ is a theorem of $\mathsf{UTB(ACA)}$. We instantiate the quantifiers in (9.10) and rename variables to get:

$$Trsl(\vec{y}) \to \forall x(T\underset{.}{s}(\ulcorner\tau(\varphi)(x, \dot{\vec{z}}, \dot{\vec{y}})\urcorner, x) \leftrightarrow \tau(\varphi)(x, \vec{z}, \vec{y}))$$

Since PA proves that $\ulcorner\tau(\varphi)(x, \dot{\vec{z}}, \dot{\vec{y}})\urcorner$ is the code of a formula of $\mathcal{L}_T$ that has exactly x free, we get:

$$Trsl(\vec{y}) \to (Fm_T^1(\ulcorner\tau(\varphi)(x, \dot{\vec{z}}, \dot{\vec{y}})\urcorner) \wedge \forall x(T\underset{.}{s}(\ulcorner\tau(\varphi)(x, \dot{\vec{z}}, \dot{\vec{y}})\urcorner, x) \leftrightarrow \varphi(x, \vec{z}, \vec{y})))$$

We have, provably in PA,

$$Trsl(\vec{y}) \to \underset{\cdot}{\tau}\ulcorner(\varphi)(x,\dot{\vec{z}},\tau^{-1}\dot{\vec{y}})\urcorner = \ulcorner\tau(\varphi)(x,\dot{\vec{z}},\dot{\vec{y}})\urcorner$$

This implies:

$$Trsl(\vec{y}) \wedge (Trsl^{0}(\ulcorner\tau(\varphi)(x,\dot{\vec{z}},\dot{\vec{y}})\urcorner) \to \forall x(T\underset{\cdot}{s}(\ulcorner\tau(\varphi)(x,\dot{\vec{z}},\dot{\vec{y}})\urcorner,x) \leftrightarrow \varphi(x,\vec{z},\vec{y}))))$$

and, by logic:

$$Trsl(\vec{y}) \to \exists v(Trsl(v) \wedge \forall x(T\underset{\cdot}{s}(v,x) \leftrightarrow \varphi^{*}(x,\vec{z},\vec{y})))$$

Now we re-introduce universal quantifiers:

$$\forall\vec{z}\forall\vec{y}((Trsl(\vec{y}) \to \exists v(Trsl(v) \to \forall x(T\underset{\cdot}{s}(v,x) \leftrightarrow \varphi^{*}(x,\vec{z},\vec{y}))))$$

This is the translation of the comprehension axiom for φ. □

Definition 9.4.5. For $S \subseteq \omega$ and $\varphi(x_0) \in \mathcal{L}_2^{+}$, let $S_{\#\tau(\varphi)} = \{n|\ \#\tau(\varphi)(\bar{n}) \in S\}$. Let $\mathcal{M}_S = \{S_{\#\tau(\varphi)}|\varphi \in \mathcal{L}_2^{+}\}$. Then $(\mathbb{N},\mathcal{M}_S)$ is a structure for $\mathcal{L}_2^{+}$, where the constants $\overline{S_\varphi}$ are interpreted by $S_{\#\tau(\varphi)}$.

If h is a variable assignment for $(\mathbb{N},\mathcal{M}_S)$, define the assignment h^{*} for $(\mathbb{N},S)$ by $h^{*}(x_{2i}) = h(x_i)$ and $h^{*}(x_{2i+1}) = \min\{\#\tau(\varphi)\ |S_{\#\tau(\varphi)} = h(X_i)\}$.

As before, we can prove:

Proposition 9.4.6. *Let $S \subseteq \omega$, let $\varphi(x,y_1,\ldots,y_r,X_1,\ldots,X_l) \in \mathcal{L}_2^{+}$, and let h be an assignment for $(\mathbb{N},\mathcal{M}_S)$. Then:*

$$(\mathbb{N},\mathcal{M}_S),h \vDash \varphi \Leftrightarrow (\mathbb{N},S),h^{*} \vDash \tau(\varphi)$$

Proposition 9.4.7. *The theory* $\mathsf{UTB}(\mathsf{ACA})$ *has an ω-model.*

Proof. Let $X_0 = \{\tau(\varphi)|\varphi \in \mathcal{L}_{PA},\mathbb{N} \vDash \varphi\}$. Let

$$X_{n+1} = \{\tau(\varphi)|\varphi \in L_n,\varphi \text{ arithmetical, } (\mathbb{N},\mathcal{M}_{X_n}) \vDash \varphi\}$$

Finally, let $S = X_\omega = \bigcup_n X_n$. There are a couple of things to notice:

- $\mathcal{M}_{X_n} = \{X \subseteq \omega|X \text{ is } \Pi^0_\omega\}$ for all n
- $X_n \subseteq X_{n+1}$
- if $\varphi \in L_n$ is arithmetical, then $(X_{n+1})_{\#\tau(\varphi)} = \{m|(\mathbb{N},\mathcal{M}_{X_n}) \vDash \varphi(\overline{m})\}$
- if $\varphi \in L_n$, then $(X_k)_{\#\tau(\varphi)} = (X_m)_{\#\tau(\varphi)}$ for all $n \leqslant k \leqslant m$

Hence, for every arithmetical φ and natural number n,

$$(\mathbb{N},\mathcal{M}_{X_\omega}) \vDash \overline{n} \in \overline{S_{\#\tau(\varphi)}} \leftrightarrow \varphi(\overline{n})$$

whence by the translation lemma and standardness $(\mathbb{N},X_\omega) \vDash \mathsf{UTB}(\mathsf{ACA})$. □

Part IV.

Appendix

10. Ordinal notations

The truth predicates of the Tarskian hierarchy are not indexed by ordinals themselves but rather by ordinal notations, that is by some subset of natural numbers equipped with some recursive ordering of suitable order type. This section surveys the relevant background material on ordinal notations needed for this book. In our exposition, we largely follow chapter 3 of Pohlers [66].

Definition 10.0.8. If $X \subseteq ON$ is a class of ordinals, we denote by en_X its enumerating function, i.e. the function which enumerates the members of X in order of increasing magnitude.

Definition 10.0.9. The class $\mathbb{H}$ of *principal* or *additively indecomposable* ordinals is defined as follows. $\mathbb{H} := \{\alpha \in ON | \alpha \neq 0 \wedge \forall \zeta, \eta < \alpha(\zeta + \eta < \alpha)\}$.

Proposition 10.0.10. *The principal ordinals are exactly those of the form ω^α. Thus: $\omega^\alpha = en_{\mathbb{H}}(\alpha)$.*

$\omega^0 = 1, \omega^1 = \omega, \omega^2 = \omega \cdot \omega.$

Definition 10.0.11. α is an ϵ-number iff $\alpha = \omega^\alpha$. The least such ordinal is called ϵ_0.

Thus ϵ_0 is the least fixed point of the function $en_{\mathbb{H}}$.

Proposition 10.0.12 (Cantor Normal Form)**.** *For every $\alpha \neq 0$ there are uniquely determined ordinals $\beta_1 \geqslant \ldots \geqslant \beta_n$ such that $\alpha = \omega^{\beta_1} + \ldots + \omega^{\beta_n}$.*

Proof. If $\alpha \in \mathbb{H}$, then $\alpha = \omega^\beta$ for some $\beta \leqslant \alpha$ and we are done. If $\alpha \notin \mathbb{H}$, then there are $\zeta, \eta < \alpha$ with $\alpha = \zeta + \eta$. By I.H. we have $\zeta = \omega^{beta_1} + \ldots + \omega^{\beta_n}$ and $\eta = \omega^{\gamma_1} + \ldots + \omega^{\gamma_m}$. Let $j \leqslant n$ be maximal such that $\beta_j \geqslant \gamma_1$. Then $\alpha = \omega^{\beta_1} + \ldots + \omega^{\beta_j} + \omega^{\gamma_1} + \ldots + \omega^{\gamma_m}$. □

Proposition 10.0.13. *If $\alpha = \omega^{\beta_1} + \ldots + \omega^{\beta_m}$ and $\gamma = \omega^{\zeta_1} + \ldots + \omega^{\zeta_n}$, then $\alpha < \gamma$ iff one of the following conditions hold:*

1. *$m < n$ and $\beta_i = \zeta_i$ for all $i \leqslant m$,*

2. *there is $j < m$ such that $\beta_i = \zeta_i$ for all $i < j$ and $\beta_j < \zeta_j$.*

10.1. Notation for ordinals below ϵ_0

It follows from Cantor's Normal Form theorem that every ordinal $\alpha < \epsilon_0$ can be written in the form $\omega^{\beta_1} + \ldots + \omega^{\beta_n}$ with $\alpha > \beta_1 > \ldots > \beta_n$. This opens the possibility of a notation system (Gödelcoding) for ordinals $< \epsilon_0$.

Definition 10.1.1. We simultaneously define the set $\mathrm{OT}_1 \subseteq \omega$ and the function $|| : OT_1 \to ON$ as follows.
$0 \in \mathrm{OT}_1$ and $|0| = 0$.
If $a_1, \ldots, a_n \in \mathrm{OT}_1$ and $|a_1| \geqslant \ldots \geqslant |a_n|$ then $(a_1, \ldots, a_n) \in \mathrm{OT}_1$ and $|(a_1, \ldots, a_n)| = \omega^{|a_1|} + \ldots + \omega^{|a_n|}$.
We set $a \prec_1 b$ iff ($a, b \in OT_1$ and $|a| < |b|$).

In the above definition, $(a_1, \ldots, a_n)$ is a code for the sequence $a_1 \ldots a_n$ (using prime numbers or Cantor's pairing function etc).

Proposition 10.1.2. *Both OT_1 and the relation $\prec_1$ are primitive recursive.*

Proof. By simultaneous course-of-value recursion:
$x \in OT_1 \leftrightarrow Seq(x) \wedge (x = 0 \vee \forall i < lh(x)((x)_i \in OT_1 \wedge (i + 1 < lh(x) \to (x)_{i+1} \preccurlyeq_1 (x)_i)))$, and
$x \prec_1 y \leftrightarrow x, y \in OT_1 \wedge ((x = 0 \wedge y \neq 0) \vee (lh(y) = 1 \wedge x \preccurlyeq_1 (y)_0) \vee$
$\exists j < min\{lh(x), lh(y)\} \forall i < j((x)_i = (y)_i \wedge (x)_j \prec_1 (y)_1) \vee$
$(lh(x) < lh(y) \wedge (\forall i < lh(x)((x)_i = (y)_i)))$. □

Proposition 10.1.3. *Every notation $a \in OT_1$ denotes an ordinal $< \epsilon_0$. Conversely, every ordinal $< \epsilon_0$ has a notation $a \in OT_1$.*

Proof. By induction. □

Example 10.1.4. *We have $0 \in OT_1$ and $|0| = 0$, $(0) \in OT_1$ and $|(0)| = \omega^0 = 1$, $(0, 0) \in OT_1$ and $|(0, 0)| = \omega^0 + \omega^0 = 1 + 1 = 2$, etc. $((0)) \in OT_1$ and $|((0))| = \omega^{|(0)|} = \omega^1 = \omega$, $((0), 0) \in OT_1$ and $|((0), 0)| = \omega^{|(0)|} + \omega^{|0|} = \omega^1 + \omega^0 = \omega + 1$.*

If $|a| = \alpha$, we set $\#\alpha := a$. Notice that there is some finite ordinal n such that $\#n > \#\omega$. We sometimes identify α with its code $\#\alpha$. We write the numeral of the code of α simply as $\overline{\alpha}$ instead of $\overline{\#\alpha}$. We reserve α, β, ζ for variables ranging over ordinal terms.

Gentzen has shown that PA proves transfinite induction for every δ up to but exluding ϵ_0.

Theorem 10.1.5 (Gentzen). *For all φ and $\delta \prec \epsilon_0$:*

$$\mathsf{PA} \vdash \forall \alpha (\forall \beta \prec \alpha \varphi(\beta) \to \varphi(\alpha)) \to \forall \zeta \prec \overline{\delta} \varphi(\zeta).$$

10.2. Notation for ordinals below Γ_0

Since ϵ_0 is a fixed point of the function ω^α, Cantor's normal form does not automatically generate a gödelization of ϵ_0. To get a notation system for ordinals $\geqslant \epsilon_0$ we need a decomposition for additively indecomposable ordinals too.

If f is a function from the ordinals to the ordinals, let $Fix(f)$ denote the class of fixed-points of f. The α-critical ordinals are defined as follows.

Definition 10.2.1. Let $Cr(0) = \mathbb{H}, Cr(\alpha{+}1) = Fix(en_{Cr(\alpha)}), Cr(\lambda) = \bigcap_{\alpha<\lambda} Cr(\alpha)$. Furthermore, define the Veblen-function φ_α as $en_{Cr(\alpha)}$.

Every set $Cr(\alpha)$ is a subset of $\mathbb{H}$, and the $Cr(\alpha)$ are decreasing.

Example 10.2.2. *φ_0 is the enumerating function of the additively indecomposoable ordinals $\mathbb{H}$, that is $\varphi_0(\alpha) = en_{\mathbb{H}}(\alpha) = \omega^\alpha$. φ_1 is the enumerating function of the fixed points of φ_0, that is, the enumerating function of the ϵ-numbers, that is $\varphi_1 = en_{Cr(1)} = en_{Fix(en_{\mathbb{H}})}$ and therefore $\varphi_1(\alpha) = \epsilon_\alpha$. Notice that $\epsilon_0 = \varphi_0(\epsilon_0) = \varphi_1(0)$.*

An ordinal α is called strongly critical iff $\alpha \in Cr(\alpha)$. We denote the least strongly critical ordinal by Γ_0. This is the so-called Feferman-Schütte ordinal. Every ordinal $\alpha < \Gamma_0$ can be written in the form $\varphi_{\zeta_1}(\beta_1) + \ldots + \varphi_{\zeta_n}(\beta_n)$ with $\beta_i, \zeta_i < \alpha$.

Notice that $0 \notin Cr(0) = \mathbb{H} = \{\omega^0, \omega^1, \ldots\}$. Furthermore, $1 \notin Cr(1) = \{\epsilon_0, \epsilon_1, \ldots\}$. Thus $\epsilon_0 \notin Cr(2)$, and therefore $\Gamma_0 > \epsilon_0$.

Proposition 10.2.3. *For every additively indecomposable ordinal α that is not strongly critical we find $\beta, \gamma < \alpha$ such that $\alpha = \varphi_\beta(\gamma)$.*

The preceeding lemma, combined with Cantor's normal form theorem, provides us with a means to gödelize all ordinals $< \Gamma_0$.

Definition 10.2.4. We simultaneously define sets PT, $\mathrm{OT}_2 \subseteq \omega$ and the function $|| : OT_2 \to ON$ as follows.

$0 \in \mathrm{OT}_2$ and $|0| = 0$.

If $a_1, \ldots, a_n \in \mathrm{PT}$ and $|a_1| \geqslant \ldots \geqslant |a_n|$ then $(1, a_1, \ldots, a_n) \in \mathrm{OT}_2$ and $|(1, a_1, \ldots, a_n)| = |a_1| + \ldots + |a_n|$.

If $a_1, \ldots, a_2 \in \mathrm{OT}_2$ then $(2, a_1, a_2) \in \mathrm{PT}$ and $|(2, a_1, a_2)| = \varphi_{|a_1|}(|a_2|)$.

$\mathrm{PT} \subseteq \mathrm{OT}_2$.

We set $a \prec_2 b$ iff ($a, b \in OT_2$ and $|a| < |b|$).

Proposition 10.2.5. *Both OT_2 and the relation $\prec_2$ are primitive recursive.*

Proposition 10.2.6. *Every notation $a \in OT_2$ denotes an ordinal $< \Gamma_0$. Conversely, every ordinal $< \Gamma_0$ has a notation $a \in OT_2$.*

Example 10.2.7. *We have $0 \in OT_2$ and $|0| = 0$; hence $(2, 0, 0) \in PT$ and $|(2, 0, 0)| = \varphi_0(0) = \omega^0 = 1$, hence $(1, (2, 0, 0), (2, 0, 0)) \in OT_2$ and $|(1, (2, 0, 0), (2, 0, 0))| = \varphi_0(0) + \varphi_0(0) = 1 + 1 = 2$, etc. Notice that OT_1 and OT_2 give different codes to the ordinal 2. We have $\epsilon_0 = \varphi_1(0) = |(2, (2, 0, 0), 0)|$.*

10.3. Kleene's $\mathcal{O}$

In order to formalize the whole Tarskian hierarchy (as in chapter 8) we need an even more encompassing notation system.

Definition 10.3.1 (Kleene's $\mathcal{O}$)**.** $0 \in \mathcal{O}$ and $|0| = 0$.

If $i \in \mathcal{O}$ and $|i| = \alpha$, then $2^i \in \mathcal{O}$ and $|2^i| = \alpha + 1$ and $i <_{\mathcal{O}} 2^i$.

Suppose $\{e\}$ is the e-th partial recursive function. If e is total, with range contained in $\mathcal{O}$, and for every natural number n we have $\{e\}(n) <_{\mathcal{O}} \{e\}(n+1)$, then $3 \cdot 5^e \in \mathcal{O}$, $\{e\}(n) <_{\mathcal{O}} 3 \cdot 5^e$ for each n and $|3 \cdot 5^e| = sup_k|\{e\}(k)|$, i.e. $3 \cdot 5^e$ is a notation for the limit of the ordinals γ_k where $|\{e\}(k)| = \gamma_k$ for every natural number k.

$<_{\mathcal{O}}$ is transitive.

Definition 10.3.2. $\omega_1^{CK} := sup\{ot(\prec)| \prec \subseteq \omega \times \omega$ is primitive recursive$\}$

Proposition 10.3.3. *Every notation $a \in \mathcal{O}$ denotes an ordinal $< \omega_1^{CK}$. Conversely, every ordinal $< \omega_1^{CK}$ has a notation $a \in \mathcal{O}$.*

Proposition 10.3.4. *$\mathcal{O}$ is Π^1_1-complete.*

Proposition 10.3.5. *For any p, the set $\{q|q <_{\mathcal{O}} p\}$ is recursively enumerable, and in fact uniformely so.*

That is, there is a recursive function f such that for all $n \in \mathcal{O}$,

$$\{m|m <_{\mathcal{O}} n\} = \{m|\exists z\{f(n)\}(z) = m\},$$

where $\{f(n)\}$ denotes the recursive partial function with index $f(n)$.

Proposition 10.3.6 (Jockusch)**.** *There exists a Π^1_1 path through $\mathcal{O}$ each initial segment of which is recursive.*

Then Tarski hierarchies can be understood in the following way (due to Halbach [33]).

Definition 10.3.7.

1. Let X be a subset of ω and $\prec$ a well-ordering of X. The Tarski hierarchy over $X, \prec$ is the set of languages $\mathcal{L}_T^k$ for $k \in X$, where $\mathcal{L}_T^k$ is $\mathcal{L}_{PA}$ expanded by all truth predicates T_n for $n \prec k$.

2. A Tarski hierarchy over $X, \prec$ is called boundedly recursive iff all initial segments of $\prec$ are recursive.

3. A Tarski hierarchy over $X, \prec$ is called recursive iff $\prec$ is recursive.

Proposition 10.3.8. *There is a boundedly recursive Tarski hierarchy of height ω_1^{CK}. For every $\alpha < \omega_1^{CK}$ there is a recursive Tarski hierarchy of height α.*

Proof. The first claim follows from Proposition 10.3.6. The second is obvious. □

11. Recursion Theory

In order to give a definition of the hyperarithmetical sets, and to state certain important relationships between these and certain subsystems of analysis, we survey some concepts from recursion theory. Most definitions in this section are only given for sets, i.e. unary relations, but they can be generalized to arbitrary relations in a straightforward way.

11.1. Indices

Let f be a recursive partial function (of one argument). Let

$$Sub_n(\#\varphi(x_1, \dots, x_n), a_1, \dots, a_n) = \#\varphi(\overline{a}_1, \dots, \overline{a}_n)$$

For given f, we find a recursive total predicate R such that $G_f(x, y) \Leftrightarrow \exists w R(y, w, x)$. Let $\varphi(\vec{z})$ be a formula that defines R. Such a φ exists because R is recursive. Notice that $(a_1, a_2, a_3) \in R$ iff $\mathsf{PA} \vdash \varphi(\overline{a}_1, \overline{a}_2, \overline{a}_3)$. Let $e = \#\varphi(\vec{z})$. Hence $(a_1, a_2, a_3) \in R$ iff $\exists v Proof(v, sb_3(e, a_1, a_2, a_3))$. Thus

$$\exists w R(y, w, x) \Leftrightarrow \exists w \exists v Proof(v, Sub_3(e, y, w, x))$$

By contraction of quantifiers, the right-hand side of the above biconditional is the case iff $\exists z Proof(z_0, Sub_3(e, y, z_1, x))$.

But then $f(x) \simeq (\mu z(Proof(z_0, Sub_3(e, z_1, z_2, x))))_0$. We write

$$T_1(e, x, z) \leftrightarrow Proof(z_0, Sub_3(e, z_1, z_2, x)))$$

Thus $f(x) \simeq (\mu z(T_1(e, x, z))_0$. We say that e is an index of f and set $\{e\} = f$.

A number e is an RE-index of a set P iff $P(x) \leftrightarrow \exists z T_1(e, x, z)$. If e satisfies this equation, we set $W^1_e = P$. Thus W^1_e is the domain of the e-th recursive partial function.

11.2. The Arithmetical and the Analytical Hierarchy

Definition 11.2.1. 1. A formula $\varphi \in \mathcal{L}_{PA}$ is called Δ^0_0 iff it contains no unbounded quantifier.

2. A formula $\varphi \in \mathcal{L}_{PA}$ is $\Pi^0_n(\Sigma^0_n)$ iff its has the form $Q_1x_1 \dots Q_nx_n\psi$, where ψ is Δ^0_0, $Q_1 \dots Q_n$ is a string of alternating quantifiers, and Q_1 is universal (existential).

Definition 11.2.2 (The Arithmetical Hierarchy)**.** A set $X \subseteq \omega$ is $\Pi^0_n(\Sigma^0_n)$ iff $X = \{n|\mathbb{N} \vDash \varphi(\overline{n})\}$, where $\varphi(x)$ is a $\Pi^0_n(\Sigma^0_n)$ formula with exactly x free. A set $X \subseteq \omega$ is Δ^0_n iff X is both Π^0_n and Σ^0_n.

Proposition 11.2.3. $\Delta^0_0 = \Delta^0_1$ *=recursive sets.* Σ^0_1 *=recursively enumerable sets.*

Definition 11.2.4.

1. A formula $\varphi \in \mathcal{L}_2$ is called *arithmetical* iff it contains no second-order quantifiers. Note that such a formula might contain free second order variables.

2. A formula $\varphi \in \mathcal{L}_2$ is $\Pi^1_n(\Sigma^1_n)$ iff its has the form $Q_1X_1 \dots Q_nX_n\psi$, where ψ is arithmetical, $Q_1 \dots Q_n$ is a string of alternating second order quantifiers, and Q_1 is universal (existential).

Definition 11.2.5 (The Analytical Hierarchy)**.** A set $X \subseteq \omega$ is $\Pi^1_n(\Sigma^1_n)$ iff $X = \{n|(\mathbb{N}, \wp(\omega)) \vDash \varphi(\overline{n})\}$, where $\varphi(x)$ is a $\Pi^1_n(\Sigma^1_n)$ formula with exactly x free. A set $X \subseteq \omega$ is Δ^1_n iff X is both Π^1_n and Σ^1_n.

Definition 11.2.6. A set X is called Π^1_n*-hard* iff every Π^1_n-set Y is many-one reducible to X, i.e. there is a recursive function f such that $n \in Y$ iff $f(x) \in X$. A set X is called Π^1_n*-complete* iff X is Π^1_n-hard and X is a Π^1_1-set.

11.3. Hyperarithmetical sets

Definition 11.3.1. We inductively define a set of H-indices as follows. Recall the definition of W^1_e as the domain of the e-th partial recursive function.

- For each $e, (0, e)$ is an H-index.
- If e is an H-index, then $(1, e)$ is an H-index.
- If every number in W^1_e is an H-index, then $(2, e)$ is an H-index.

For each H-index i we define a set J_i as follows.

- If $i = (0, e)$, then $J_i = W^1_e$.
- If $i = (1, e)$, then $J_i = (\omega \setminus W^1_e)$.
- If $i = (2, e)$, then $J_i = \bigcup_{k \in W^1_e} J_k$.

Definition 11.3.2. A set $X \subseteq \omega$ is *hyperarithmetical* iff $P = J_i$ for some H-index i. We call i the H-index of P. The collection of all hyperarithmetical sets is denoted by HYP.

Thus, the hyperarithmetic sets are constructed by starting with the recursively enumerable sets and repeatedly taking complements and certain countable unions, namely unions of sets where there is a recursive enumeration of the H-index of the sets in question.

Alternatively, we might define iterated Turing jumps along Kleene's $\mathcal{O}$ by stipulating that $\varnothing^0 = \varnothing$, $\varnothing^{\alpha+1} = TJ(\varnothing^\alpha)$, and $\varnothing^\lambda = \{\langle n, i\rangle | i \in \varnothing^{\lambda_n}\}$. Then we say that a set X is hyperarithmetical iff X is Turing-reducible to $\varnothing^\alpha$ for some $\alpha < \omega_1^{CK}$.

Proposition 11.3.3. *$X \in HYP$ iff X is Δ^1_1.*

For a proof see Schoenfield [83], chapter 7.

Proposition 11.3.4. *$(\mathbb{N}, HYP)$ is the minimal ω-model of $\Delta^1_1 - \mathsf{CA}_0$.*

For a proof see Simpson [85], chapter VIII.

11.4. The Ramified Analytical Hierarchy

The language of ramified analysis has variables for numbers and sets of numbers. The variables for the latter are indexed by ordinals, which denote the order of the set. It is thus a ramified language as in Russell and Whitehead's *Principia mathematica*. Let ω_1 be the first uncountable ordinal. This language is used to describe the predicative sets (of natural numbers).

Definition 11.4.1. For $\alpha < \omega_1$ let $\mathcal{L}_2^\alpha = \mathcal{L}_{PA} \cup \{X_i^\beta | \beta < \alpha, i \in \omega\}$, where X_i^β is a unary second order predicate variable. $\mathcal{L}_2^{\omega_1}$ is the union of all $\mathcal{L}_2^\alpha$.

This is not a recursive language. But for every $\alpha < \omega_1^{CK}$ it is. An interpretation for $\mathcal{L}_2^\alpha$ is of the form $(\mathbb{N}, (A_\beta)_{\beta<\alpha})$, where $A_\beta \subseteq \wp(\omega)$, and the quantifiers $\forall X^\beta, \exists X^\beta$ range over A_β.

Definition 11.4.2 (Ramified Analytical Sets)**.** The sets $\mathcal{RA}_\alpha$ are defined as follows. $\mathcal{RA}_0$ = the collection of arithmetically definable sets. $\mathcal{RA}_{\alpha+1}$ = the set of $X \subseteq \omega$ such that there is a $\varphi \in \mathcal{L}_2^{\alpha+1}$ and φ defines X in $(\mathbb{N}, (\mathcal{RA}_\beta)_{\beta\leqslant\alpha})$ (where we assume that the variables X^β take values in $\mathcal{RA}_\beta$ for $\beta \leqslant \alpha$). At limits we take unions. The collection $\mathcal{RA}$ of ramified analytic sets is the union of all $\mathcal{RA}_\alpha$.

Notice that this hierarchy is monotone. Thus there is a point such that $\mathcal{RA}_\alpha = \mathcal{RA}_{\alpha+1}$. The least such α is called β_0. The hierarchy of ramified analytical sets is the second-order version of Gödel's constructible hierachy. We have $\mathcal{RA}_\alpha = L_\alpha \cap \wp(\omega)$, where L_α refers to the α-th level of the constructible hierarchy.

We owe to Cohen, Putnam et al. [69] the following result:

Theorem 11.4.3 (Cohen, Putnam et al.). $(\mathbb{N}, \mathcal{RA}_{\beta_0})$ *is the minimal* β*-model of* Z_2.

Here, a model $(\mathbb{N}, \mathcal{M})$ (where $\mathcal{M} \subseteq \wp(\omega)$) is called a β-model iff every Π^1_1-sentence that is true in the model is true in the standard model $(\mathbb{N}, \wp(\omega))$. We owe to Kleene the following important result:

Theorem 11.4.4 (Kleene). $\mathcal{RA}_{\omega_1^{CK}} = HYP$.

Feferman [18] gives an axiomatization RA of the ramified analytical hierarchy for levels $< \Gamma_0$. In [20] he shows that the Kripke-Feferman theory of truth KF interprets all levels $< \epsilon_0$ of RA.

11.5. Subsystems of second-order arithmetic

Definition 11.5.1. The language $\mathcal{L}_2$ of second-order arithmetic is obtained from $\mathcal{L}_{PA}$ by adding the binary relation symbol $\in$ plus *set variables* $X_0, X_1, X_2, \ldots$ (Let us call $v_0, v_1, \ldots$ *number* variables.) This gives us new formulae of the form $t \in X$ and $\forall X \varphi$. $\mathcal{L}_2$ is a two-sorted first-order language with usual (first-order) rules for both set and number quantifiers. A formula φ of $\mathcal{L}_2$ is called *arithmetical* if does not contain bound set variables. (Free set variables are allowed.)

Definition 11.5.2. Z_2 is the theory in $\mathcal{L}_2$ that contains in addition to the axioms of PA all comprehension axioms

$$\forall \vec{Y} \forall \vec{y} \exists X \forall x (x \in X \leftrightarrow \varphi(x, \vec{y}, \vec{Y})),$$

where $\varphi(x, \vec{y}, \vec{Y})$ is a formula of $\mathcal{L}_2$ with all free variables displayed, and the induction axiom

$$\forall X (0 \in X \wedge \forall x (x \in X \to x + \overline{1} \in X) \to \forall x (x \in X)).$$

Notice that Z_2 proves the induction axiom scheme

$$\varphi(\overline{0}) \wedge \forall x (\varphi(x) \to \varphi(S(x)) \to \forall x \varphi(x),$$

where $\varphi(x)$ is an $\mathcal{L}_2$-formulae possibly containing free number and set variables. Z_2 is an axiomatic first-order theory in a two-sorted language; it must not to be confused with the set of second-order sentences that are true in the standard model of second-order arithmetic, $(\mathbb{N}, \wp(\omega))$. We denote by Z_2^- the subsystem of Z_2 that is obtained by restricting the comprehension axioms to formulae that do not contain free set variables (bound set variables are allowed).

Definition 11.5.3. ACA_0 is the theory in $\mathcal{L}_2$ that contains in addition to the axioms of PA all comprehension axioms

$$\forall\vec{Y}\forall\vec{y}\forall x(\varphi(x,\vec{y},\vec{Y}) \leftrightarrow \psi(x,\vec{y},\vec{Y})) \to \forall\vec{Y}\forall\vec{y}\exists X\forall x(x \in X \leftrightarrow \varphi(x,\vec{y},\vec{Y})),$$

where $\varphi(x,\vec{y},\vec{Y}) \in \mathcal{L}_2$ is an arithmetical formula, and the induction axiom

$$\forall X(0 \in X \wedge \forall x(x \in X \to x + \overline{1} \in X) \to \forall x(x \in X)).$$

The subscript 0 indicates that induction is restricted. We denote by ACA the system obtained from ACA_0 by adding all instances of the induction axiom scheme.

Definition 11.5.4. $\Delta^1_1 - \mathsf{CA}_0$ is the theory in $\mathcal{L}_2$ that contains in addition to the axioms of PA all comprehension axioms

$$\forall\vec{Y}\forall\vec{y}\forall x(\varphi(x,\vec{y},\vec{Y}) \leftrightarrow \psi(x,\vec{y},\vec{Y})) \to \forall\vec{Y}\forall\vec{y}\exists X\forall x(x \in X \leftrightarrow \varphi(x,\vec{y},\vec{Y})),$$

where $\varphi(x,\vec{y},\vec{Y}) \in \mathcal{L}_2$ is a Π^1_1-formula and $\psi(x,\vec{y},\vec{Y}) \in \mathcal{L}_2$ is a Σ^1_1-formula, and the induction axiom

$$\forall X(0 \in X \wedge \forall x(x \in X \to x + \overline{1} \in X) \to \forall x(x \in X)).$$

Definition 11.5.5. The language of ID_1 extends the language $\mathcal{L}_{PA}$ by a predicate constant $\overline{I}_\varphi$ for every *arithmetical* $\mathcal{L}_2$-formula $\varphi(v_0, X_0)$ (with exactly the displayed variables free) in which the free set variable X_0 occurs only *positively* (i.e. it does not appear in the scope of an odd number of negation signs). We may identify expressions of the form $\overline{I_\varphi}(t)$ with $t \in \overline{I_\varphi}$ and regard $\overline{I_\varphi}$ as a set constant. On the intended interpretation, the set constant $\overline{I_\varphi}$ is interpreted by the least fixed point generated (or the inductive relation defined) by the formula φ.

Definition 11.5.6. ID_1 is the theory in $\mathcal{L}_{ID_1}$ that contains in addition to the axioms of PA and full induction in $\mathcal{L}_{ID_1}$ all axioms of the form

$$\forall x(\varphi(x, I_\varphi) \to I_\varphi(x))$$

and

$$\forall x(\varphi(x,\psi) \to \psi(x)) \to \forall x(I_\varphi(x) \to \psi(x))$$

Here, $\varphi(x,\psi)$ is obtained from $\varphi(x,X)$ by replacing every occurrence of $t \in X$ by $\psi(t)$ and of $\neg(t \in X)$ by $\neg\psi(t)$. The system $\widehat{\mathsf{ID}}_1$ is the theory in $\mathcal{L}_{ID_1}$ that contains in addition to the axioms of PA and full induction in $\mathcal{L}_{ID_1}$ all axioms of the form

$$\forall x(\varphi(x, I_\varphi) \leftrightarrow I_\varphi(x))$$

The proof-theoretic ordinal of ID_1 is the Bachmann-Howard ordinal. Cf. Pohlers [66, ch. 9]. The Bachmann-Howard ordinal is also the proof-theoretic ordinal of the system $\Pi^1_1 - \mathsf{CA}^-_0$.

12. Graph theory

A *directed graph* (short:a *digraph* or simply a *graph*) G consists of a set $V(G)$, the *vertices* (or *nodes*) of G, and of a set $A(G)$ of ordered pairs of vertices, called *arcs* of G. If $x, y \in V(G)$ we denote an arc from x to y by (x, y); we call x its *tail* and y its *head.* For any vertex x call y an *out-neighbour* of x iff $(x, y) \in A(G)$ and an *in-neighbour* iff $(y, x) \in A(G)$. A graph H is a *subgraph* of G iff $V(H) \subseteq V(G)$ and $A(H) \subseteq A(G)$. In this case we also say that G *contains* H and write $H \subseteq G$.

A non-empty graph P (i.e. a graph with at least one vertex) is called a *path* (from a to b, of length $n-1$) iff there is an enumeration $(v_0, v_2, ..., v_n)$ of $V(P)$ such that for all $0 \leq i, j \leq n$ $(v_i, v_j) \in A(P)$ iff $j = i+1$ with $a = v_0$ and $b = v_n$. Note that a graph with one vertex and no arcs is a path of length 0. We call such a path *trivial.* A graph D is called a *double-path* (from a to b) iff there are non-trivial pathes P_1, P_2 from a to b such that $V(P_1) \cap V(P_2) = \{a, b\}$ and $V(D) = V(P_1) \cup V(P_2)$ and $A(D) = A(P_1) \cup A(P_2)$. A graph C is called a *cycle* (of length $n+1$) iff there is a (possibly trivial) path P of length n from a to b such that $V(C) = V(P)$ and $A(C) = A(P) \cup \{(b, a)\}$. A cycle of length 1 is called a *loop.*

For any graph G, call an infinite sequence of vertices $(v_0, v_2, ...)$ of $V(G)$ an *infinite walk* in G iff for all $i \in \omega (v_i, v_{i+1}) \in A(G)$. Analogously we can define a finite walk. Note that one and the same vertex may occur more than once in a walk, while the sequence enumerating the vertices of a path P contains, by definition, every vertex of P only once. A graph G is called *well-founded* iff there is no infinite walk in G.

We call a graph H a *subdivision* of G iff is the result of replacing each $(x, y) \in A(G)$ by some path from x to y (possibly of length 1).

Call a graph G *strongly connected* iff any two distinct vertices x, y of G are joint by a path $P \subseteq G$ from x to y or from y to x. A graph is a *tree* iff it is strongly connected and contains no cycle.

A graph H is an *induced subgraph* of a graph G iff H is a subgraph of G and each arc of G between two vertices of H is also an arc of H. In this case we also say that $V(H)$ *spans* H in G and write $H = G[V(H)]$.

13. Bibliography

[1] Armour-Garb, B. Challenges to deflationary theories of truth. *Philosophy Compass 7*, 256-266 (2012).

[2] Armour-Garb, B., and Beall, J. C., Eds. *Deflationism and Paradox.* Oxford University Press, New York, 2005.

[3] Beall, J. C. Is Yablo's Paradox non-circular? *Analysis 61* (2001), 176—187.

[4] Beall, J. C. Transparent disquotationalism. In *Deflationism and Paradox*, J. C. Beall and B. Armour-Garb, Eds. Oxford University Press, 2005, pp. 7–22.

[5] Beall, J. C., Ed. *Revenge of the liar. New essays on the paradox.* Oxford University Press, 2007.

[6] Beall, J. C. *Spandrels of Truth.* Oxford University Press, Oxford, 2009.

[7] Beringer, T., and Schindler, T. Reference-graphs, games for truth, and paradox. unpublished manuscript, 2015.

[8] Burgess, J. P. The truth is never simple. *Journal of Symbolic Logic 51* (1986), 663–681.

[9] Cain, J., and Damnjanovic, Z. On the Weak Kleene scheme in Kripke's theory of truth. *Journal of Symbolic Logic 56* (1991), 1452–1468.

[10] Cantini, A. Notes on formal theories of truth. *Zeitschr. f. math. Logik und Grundlagen d. Math. 35* (1989), 97–130.

[11] Cantini, A. A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic 55* (1990), 244–259.

[12] Carnap, R. *Logical foundations of probability.* University of Chicago Press, Chicago, 1950.

[13] Cobreros, P., Egré, P., Ripley, D., and van Rooij, R. Reaching Transparent Truth. *Mind 122* (2013), 841–866.

[14] Cook, R. T. Patterns of paradox. *Journal of Symbolic Logic 69*, 3 (2004), 767–774.

[15] Cook, R. T. There are non-circular paradoxes (but Yablo's Isn't One of Them!). *The Monist 89*, 1 (2006), 118–149.

[16] Craig, W., and Vaught, R. Finite axiomatizability using additional predicates. *Journal of Symbolic Logic 23* (1958), 289–308.

[17] Esser, O. On the consistency of a positive theory. *Mathematical Logic Quaterly 45* (1999), 105–116.

[18] Feferman, S. Systems of predicative analysis. *Journal of Symbolic Logic 29* (1964), 1–30.

[19] Feferman, S. Towards useful type-free theories, I. *Journal of Symbolic Logic 49* (1984), 75–111.

[20] Feferman, S. Reflecting on incompleteness. *Journal of Symbolic Logic 56* (1991), 1–49.

[21] Feferman, S. Axioms for determinateness and truth. *Review of Symbolic Logic 1* (2008), 204–217.

[22] Field, H. Deflationist views of meaning and content. *Mind 103* (1994), 249–284.

[23] Field, H. Deflating the Conservativeness Argument. *Journal of Philosophy 96* (1999), 533–540.

[24] Field, H. A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic 32* (2003), 139–177.

[25] Field, H. Solving the paradoxes, escaping revenge. In *Revenge of the liar*, J. Beall, Ed. Oxford University Press, 2007, pp. 78–144.

[26] Field, H. *Saving truth from paradox.* Oxford University Press, New York, 2008.

[27] Frege, G. The thought: a logical inquiry. *Mind*, 65 (1956), 289–311.

[28] Friedman, H., and Sheard, M. An axiomatic approach to self–referential truth. *Annals of Pure and Applied Logic 33* (1987), 1–21.

[29] Fujimoto, K. Relative truth definability of axiomatic truth theories. *Bulletin of Symbolic Logic 16* (2010), 305–344.

[30] Gödel, K. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems.* Basic Books, New York, 1962.

[31] Gupta, A. Truth and paradox. *Journal of Philosphical Logic 11* (1982), 1–60.

[32] Gupta, A., and Belnap, N. D. *The Revision Theory of Truth.* MIT Press, Cambridge, 1993.

[33] Halbach, V. Tarski hierarchies. *Erkenntnis 43* (1995), 339–367.

[34] Halbach, V. *Axiomatische Wahrheitstheorien.* Logica Nova. Akademie Verlag, Berlin, 1996.

[35] Halbach, V. Tarskian and Kripkean truth. *Journal of Philosophical Logic 26* (1997), 69–80.

[36] Halbach, V. Disquotationalism and infinite conjunctions. *Mind 108* (1999), 1–22.

[37] Halbach, V. Disquotational truth and analyticity. *Journal of Symbolic Logic 66* (2001), 1959–1973.

[38] Halbach, V. *Axiomatic Theories of Truth.* Cambridge University Press, Cambridge, 2011.

[39] Halbach, V., and Horsten, L. The deflationist's axioms for truth. In *Deflationism and Paradox*, B. Armour-Garb and J. C. Beall, Eds. Oxford University Press, 2005.

[40] Heck, R. J. Self-reference and the Languages of Arithmetic. *Philosophia Mathematica III*, 15 (2007), 1–29.

[41] Herzberger, H. Paradoxes of grounding in semantics. *Journal of Philosophy 67* (1970), 145–167.

[42] Herzberger, H. Notes on naive semantics. *Journal of Philosphical Logic 11* (1982), 61–102.

[43] Hintikka, J. *Knowledge and belief. An introduction to the logic of the two notions.* Cornell University Press, 1962.

[44] Horsten, L. The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In *The many problems of realism*, P. Cortois, Ed., vol. 3 of *Studies in the general philosophy of science.* Tilburg University Press, 1995, pp. 173–187.

[45] Horsten, L. *The Tarskian Turn: Deflationism and Axiomatic Truth.* MIT Press, Cambridge, 2011.

[46] Horwich, P. *Truth*, 2nd ed. Basil Blackwell, Oxford, 1998.

[47] Horwich, P. A minimalist critique of tarski on truth. In *Deflationism and Paradox*, J. C. Beall and B. Armour-Garb, Eds. Oxford University Press, 2005, pp. 75–84.

[48] Jeroslow, R. G. Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. *Journal of Symbolic Logic 38* (1973), 359–367.

[49] Ketland, J. Deflationism and tarski's paradise. *Mind 108* (1999), 69–94.

[50] Kleene, S. C. Finite axiomatizability of theories in the predicate calculus using additional predicate symbols. *Memoirs of the american mathematical society 10* (1952), 27–68.

[51] Kleene, S. C. Quantification of number-theoretic functions. *Compositio Mathematica 14* (1959), 23–40.

[52] Kremer, M. Kripke and the Logic of Truth. *Journal of Philosophical Logic 17* (1988), 225–278.

[53] Kripke, S. Outline of a theory of truth. *Journal of Philosphy 72* (1975), 690–716.

[54] Landini, G. Russell's substitutional theory of classes and relations. In *The Cambridge Companion to Bertrand Russell*, N. Griffin, Ed. Cambridge University Press, 2003, pp. 241–286.

[55] Leitgeb, H. What Truth Depends On. *Journal of Philosphical Logic 34* (2005), 155–192.

[56] Martin, D. A. Revision and its rivals. *Philosophical Issues 8* (1997), 407–418.

[57] Maudlin, T. *Truth and Paradox.* Oxford University Press, New York, 2004.

[58] Maudlin, T. Reducing revenge to discomfort. In *Revenge of the liar. New essays on the paradox*, J. C. Beall, Ed. Oxford University Press, 2007, pp. 184–224.

[59] McGee, V. Maximal consistent sets of instances of Tarski's schema. *Journal of Philosphical Logic 21* (1992), 235–241.

[60] Meadows, T. Truth, dependence, and supervaluation: Living with the ghost. *Journal of Philosophical Logic 42* (2013), 221–240.

[61] Montagna, F., and Mancini, A. A minimal predicative set theory. *Notre Dame Journal of Formal Logic 35* (1994), 186–203.

[62] Moschovakis, Y. N. *Elementary induction on abstract structures.* Dover Publications, 1974.

[63] Paoli, F. *Substructural Logics: A Primer.* Kluwer, Dordrecht, 2002.

[64] Picollo, L. The Old-Fashioned Yablo Paradox. *Análisis Filosófico 32*, 1 (2012), 21–29.

[65] Picollo, L., and Schindler, T. Disquotation and infinite conjunctions. unpublished manuscript, 2015.

[66] Pohlers, W. *Proof Theory: The First Step into Impredicativity.* Springer, Berlin Heidelberg, 2009.

[67] Priest, G. Yablo's Paradox. *Analysis 57* (1997), 236–242.

[68] Priest, G. *In Contradiction.* Oxford University Press, New York, 2006.

[69] Putnam, H., Boyd, R., and Hensel, G. A recursion theoretic characteriaztion of the ramified analytical hierarchy. *Transactions of the American Mathematical Society 141/142* (1969), pp. 37–62.

[70] Quine, W. V. O. Two dogmas of empiricism. In *From a logical point of view*, second ed. Harvard University Press, 1961, pp. 20–46.

[71] Quine, W. V. O. *Set theory and its logic.* Harvard University Press, 1964.

[72] Quine, W. V. O. *Philosophy of Logic.* Harvard University Press, 1970.

[73] Quine, W. V. O. New foundations for mathematical logic. In *From a logical point of view.* Harvard University Press, 1980, pp. 80–101.

[74] Quine, W. V. O. On what there is. In *From a logical point of view.* Harvard University Press, 1980, pp. 1–19.

[75] Rabern, L., Rabern, B., and Macauley, M. Dangerous reference graphs and semantic paradoxes. *Journal of Philosophical Logic 42*, 5 (2013), 727–765.

[76] Ramsey, F. P. Facts and propositions. *Proceedings of the Aristotelian Society 7* (1927), 153–170.

[77] Restall, G. Minimalists about truth can (and should) be epistemicists, and it helps if they are revision theorists too. In *Deflationism and Paradox*, J. C. Beall and B. Armour-Garb, Eds. Oxford University Press, 2005, pp. 97–106.

[78] Russell, B. On denoting. *Mind 14* (1903), 479–493.

[79] Russell, B. On some difficulties in the theory of transfinite numbers and order types. In *Essays in analysis*. Allen and Unwin, London, 1973, pp. 135–164.

[80] Russell, B. On the substitutional theory of classes and relations. In *Essays in analysis*. Allen and Unwin, 1973, pp. 165–189.

[81] Schindler, T. Axioms for grounded truth. *Review of Symbolic Logic 7* (2014), 73–83.

[82] Schindler, T. A disquotational theory of truth as strong as Z_2^-. *Journal of Philosophical Logic* (2014), Online First, DOI 10.1007/s10992–014–9327–5, 18 pp.

[83] Schoenfield, J. R. *Mathematical Logic*. Addison-Wesley, 1967.

[84] Shapiro, S. Proof and Truth: Through Thick and Thin. *Journal of Philosophy 95* (1998), 493–521.

[85] Simpson, S. G. *Subsystems of Second Order Arithmetic*, Second ed. Cambridge University Press, Cambridge, 2009.

[86] Soames, S. *Understanding truth*. Princeton University Press, 1999.

[87] Sorensen, R. A. Yablo's Paradox and kindred infinite liars. *Mind 107* (1998), 137–155.

[88] Takeuti, G. *Proof Theory*, second ed. North Holland, Amsterdam, 1987.

[89] Tarski, A. The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*. Clarendon Press, Oxford, 1935, pp. 152—278.

[90] Vugt, F., and Bonnay, D. What makes a sentence be about the world? towards a unified account of groundedness. unpublished manuscript, 2009.

[91] Weir, A. Naive truth and sophisticated logic. In *Deflationism and Paradox*, J. C. Beall and B. Armour-Garb, Eds. Oxford University Press, 2005, pp. 218–249.

[92] Welch, P. Ultimate truth vis á vis stable truth. *Review of Symbolic Logic* (2008), 126–142.

[93] Welch, P. Games for truth. *Bulletin of Symbolic Logic 15*, 4 (2009), 410–427.

[94] Welch, P. The complexity of the dependence operator. *Journal of Philosophical Logic* (2014), Online First, DOI 10.1007/s10992–014–9324–8, 4 pgs.

[95] Yablo, S. Grounding, dependence, and paradox. *Journal of Philosophical Logic 11*, 1 (1982), 117–137.

[96] Yablo, S. Paradox without self-reference. *Analysis 53* (1993), 251–252.

Index