Università degli Studi di Firenze

FACOLTÁ DI LETTERE E FILOSOFIA
FACOLTÁ DI SCIENZE MATEMATICHE, FISICHE E NATURALI.
Corso di Laurea Magistrale in Logica, Filosofia e Storia della Scienza

Tesi di Laurea in Logica

Truth and Reduction.

A Case Study in Formal Philosophy.

Relatore:
Prof. Andrea Cantini

Correlatore:
Prof. Pierluigi Minari

Candidato:
Rossella Marrano

Anno Accademico 2011-2012

Doubt thou the stars are fire,
Doubt that the sun doth move,
Doubt truth to be a liar,
But never doubt I love.

William Shakespeare, *Hamlet Act II scene II*.

Contents

Introduction

This work is concerned with the notion of truth and some related logico-philosophical issues. Truth has been one of the central topics of discussion throughout the history of philosophy. The modern dispute about truth takes place on several levels and involves various philosophical disciplines: ethics, theology, metaphysics, hermeneutics, logic, epistemology, linguistics, theory of knowledge. My reflection fits into a specific school of thought which favours a logical-linguistic approach to truth. I argue that it is essential for discussions concerning broad problems (like truth, necessity, meaning, etc.) trying to make the concepts used as clear as possible. This must be done with every means: definitions, axiomatic procedures and so on. Implicitly, I commit myself with the idea of *formal philosophy*: the idea of investigating philosophical problems by means of rigorous methods of mathematical logic. Although the approach is formal, the starting point and the destination are purely philosophical issues and moreover a philosophical attitude permeates the whole research. The overall aim is to let philosophical *desiderata* communicate with formal results.

The viability of an axiomatic approach to truth has been thoroughly explored with fruitful and interesting results and many different axiom systems for truth have been discussed in the literature. One of the possible ways to investigate their respective properties is comparing them from various points of view. In this study I plan to focus on the features of a meta- and inter-theoretical inquiry in axiomatic truth theories. The aim is to emphasize the *relevance* and the effects of this analysis, since in exclusively technical papers this point is often left unexpressed, though not ignored. This will be done by analysing motivations, methods, possible outcomes and trying to let the issue benefit from a purely logical research.

The work is broadly dived into three main parts: the first part is a sort of introduction to the vast world of axiomatic theories of truth, then I shall focus on the main problem, that is why and how comparing theories of truth, and I shall introduce some hypothesis that will be sustained in the last part by presenting a case study. These parts correspond to three chapter, let us briefly expose contents and purposes:

1. The first chapter is dedicated to some axiomatic theories of truth that

will be useful throughout the work. This introduction is embodied in a more general framework: a short account of the Tarskian turn will allow us to present axiomatic truth theories as the results of a logico-philosophical strategy in dealing with truth. The underlying methodology is not an historical survey, rather the approach is synoptic: my choice is to present theories following as *leitmotiv* the efforts put in place in order to avoid paradoxes. I plan to present axiomatic theories of truth as possible solutions for semantical paradoxes. A large variety of axiomatic theories of truth have been proposed in the literature, I shall focus just on some of them.

2. The second chapter is concerned with the theme of reduction and it has a philosophico-methodological character. The starting point is the reduction between formal systems, a wide subject that will be tackled by outlining some broadly distinctions as guidelines. Once the kind of reduction I are interested in is identified, I shall address a formal overview on the most used notions of reduction. The core of the chapter is an attempt to relate the issue of reduction with truth theories. This issue will be articulated from different points of view: why comparing truth theories? how to compare them? what do we aspect from this kind of analysis? I shall argue in favour of the philosophical relevance of such an inquiry by stressing to what extent axiomatic theories of truth reveal some peculiarities which allows us to cast light on issues concerning both truth and reduction *per se*.

3. In the third chapter I shall introduce a case study in order to support the previously stated theoretical claims. A recent type-free theory of truth, Feferman's theory of determinate truth DT, will be introduced emphasizing both philosophical motivations and critical assessment. I shall report, with proof, some results about the proof theory of DT in order to underline its relationships with other theories of truth and with a mathematical system; that is to say DT will be submitted to a metatheoretical inquiry. Moreover, in the last part of the work I shall carry out an inedit comparison between this theory and a theory of truth and propositions proposed by Cantini. In this respect, I underline some obstacles to a possible reduction of one theory to the other, in spite of this such a comparison will allow me to stress some interesting philosophical points. Although the approach used in this chapter is mainly logical, logical results will be glossed with philosophical remarks.

Chapter 1

Axiomatic theories of truth

I have characterized axiomatic theories of truth as the results of a logical approach to truth. The first section is devoted to explain this claim. Then I shall describe how axiomatic theories of truth are effectively built.

1.1 A formal approach to truth

The starting-point is the word 'true' in everyday language and its main use as a grammatical predicate which takes nouns and phrases as subject:

This is true.

It is true that a word to the wise is sufficient.

The second sentence of this list is true.

What you said yesterday is true.

'True' seems to be a predicate which attributes a property — truth — to some entities. There is a considerable literature devoted to the problems whether truth is a property and, if so, in relation to what it is a property. I shall follow many contemporary philosophers' answers to these questions, which consider *truth as a property of sentences*. Truth can be considered a property since the unary predicate 'is true', the truth predicate, has a non trivial extension, i.e. an extension which is not empty nor does it consist of everything. What kind of entities populate its extension? What sort of things are true (or false)? I choose sentences (linguistic items within a specific language) as the primary bearers of truth between the different candidates: utterances, beliefs, propositions and sentences; but it should be pointed out that the sentences we consider are fully interpreted or meaningful sentences. Moreover, we assume the sentences in question to be 'eternal' sentences, i.e. sentences which are not context-dependent and, so, do not change their content according to across occasion of use.

Talking about truth, we implicitly accept a non trivial assumption: the notion of truth is an essential one, it plays an important role not only in natural languages, but also in philosophical, logical, mathematical, empirical theories. We even assume that something meaningful can be said about it. As Soames[1] observed, there are several forms of *truth skepticism*, i.e. lines of thinking that question these assumptions. Among them we can recall the beliefs that true is undefinable, unknowable, irreducibly metaphysical and that truth predicate is trivial and dispensable. I refer to his work for a deeper discussion on this topic, I just say something about the supposed triviality of truth predicate. The problem is whether we need it as independent predicate: which functions can it perform? Is it dispensable? It is often said that truth is a *disquotational device*[2], which allows us to transform a sentence into a term and *vice versa* simply by placing or removing quotation marks:

'Snow is white' is true if and only if snow is white.

Sentences like this, called *T(arski)-sentences* or *disquotation sentences*, reflect our intuitive answer to the question: when is a sentence like 'Snow is white' true? The set of such sentences is generated by collecting the instances of an axiom schema:

The sentence 'ϕ' is true if and only if ϕ.

These biconditionals connect a sentence (ϕ) with an object ('ϕ' is the name of a sentence; so it is a term, no longer a sentence). In other words, we can state a sentence turning it into an object and asserting that this object is true. This link is made possible by the truth predicate. The problem now is whether we can always eliminate the truth predicate by erasing it and removing the quotation marks. If this were the case, then the concept of truth would be *redundant*. According to some philosophers the predicate *is true* is not used to describe anything: saying "'ϕ' is true" is just a redundant way of saying ϕ. As we have just seen, this may be true to some extent as long as truth is attributed to single, explicit sentences. However, the main problem for this kind of redundancy theory is posed by examples in which the sentences that are said to be true are not directly displayed (called *blind ascriptions*):

1. Everything Aldo said yesterday is true.
2. Something Barbara believes is true.
3. Tarski's theorem is true.

[1] Cfr. Soames [44].

[2] I skip the question whether truth is *nothing more* than a device of disquotation, as diquotationalists hold.

In such cases the truth predicate cannot simply be erased, because we do not know exactly the sentences of which truth is predicated. There is a second way in which the truth predicate is essential: it allows us to express infinitely long disjunctions and conjunctions. Quine gives an example of how we can reduce 'infinite lots of sentences' to single sentences containing the truth predicate. Consider the infinite conjunction of sentences like:

If time flies then time flies.
If snow is white then snow is white.
⋮

If we want to state all these sentences by using just one sentence and without quantifying over sentences we can say:

All sentences of the form 'If p then p' are true.

The truth predicate allows us to assert in a single sentence the infinite conjunction of all sentences having that logical form. In other words, by using the truth predicate we are able to say something we could not say without it. Hence, it has an expressive power which extends the strength of our language and it must be considered as a genuine, non redundant, predicate.

Having made clear to what extent an inquiry about truth is legitimate, let us turn our attention towards the philosophical debate about it. I just want to provide a very brief account of this issue, sketching the principal trends and the turning point marked by Alfred Tarski.

The traditional philosophical debate about truth focused on substantial definitions, i.e. attempts to answer the age-old question: what is (the essence of) truth? Different proposals have been made in this direction; the most significant are the correspondence, coherence and pragmatist theories of truth. They share an important presupposition: the required definition has to be explicit, namely one which allows a complete elimination of the defined notion. These theories are distinguished by their relative views about the nature of truth and about the truth bearers. In a nutshell: the basic idea of the first is that what we say (or believe) is true if it corresponds to the facts (the way things actually are); coherence theorists hold instead that a belief is true if it is part of a coherent system of beliefs and, lastly, according to pragmatists there is a deep connection between truth and usefulness[3]. Even this rough presentation brings out that the main difficulty of definitional theories of truth is their vagueness. The *definiens* (correspondence, coherence, utility) is no clearer than the *definiendum*, that is, the notion of truth. The vagueness cannot be erased until an uncontroversial explanation of the meaning of the terms 'correspondence', 'fact', 'coherent systematic whole', etc. has been given.

[3] In each theory the notion of truth is part of a more extensive metaphysics or epistemology.

Traditional attempts to define truth embody an intuitive knowledge of the concept of truth, knowledge that everyone possesses to a certain degree; but from the point of view of formal correctness and clarity they do not seem appropriate. So, they should be considered more as explanations of some characteristics of the intuitive notion 'truth' rather than as rigorous definitions. Therefore, it would be better making our intuitive assumptions about truth explicit, without resorting to other notions that are less clear. The fist step in this direction is the reformulation of the problems in a formal setting.

It was the Polish logician, mathematician and philosopher Alfred Tarski[4] who upset the terms of the debate, starting the modern discussion about truth and articulating a theory of truth which describes the functioning of the concept of truth, and no longer its essence. Another radical shift is due to him: the reformulation of the issue in a formal framework. He pointed out the lack of formal precision which the previous theories suffer from and restricted his attention mainly to formalized languages, namely languages whose structure has been exactly specified. It means that one must characterize unambiguously the class of the expressions which are to be considered *meaningful*, defining inductively the classes of terms and sentences, providing the axioms (primitive sentences asserted without proof) and the rules of inference to deduce new sentences (theorems)[5]. Only in such languages the problem of definition of truth obtains a precise meaning and can be tackled in a rigorous way.

Having made the problem clear in a formal language, that is a fragment of English containing T, the question becomes: is it possible to define truth within it? I shall show that Tarski himself provides a negative answer: the assumption that there is a definition of truth within a given language for the same language leads to a contradiction. Therefore, the only way is take truth as a primitive notion and make clear what is expected from it. In formal terms, it means to expand the language by a new primitive predicate for truth, T, and to lay down the axioms for it. This approach does not presuppose definability, but at the same time it might be compatible with the view that truth is definable: it simply does not commit itself to the question whether truth is definable or not.

Once we have chosen non-definitional theories of truth, we are again at a crossroads: theories of truth can be further divided into two classes. There are *semantic* theories of truth and *axiomatic* ones; but, as we shall see soon, these approaches are in some ways complementary. Roughly speaking, the former are mainly interested in describing models for a language that contain the truth predicate by providing an interpretation of T (model-theoretic ap-

[4] See Tarski [45]

[5] That which he call *formalized language* in modern terminology is a *deductive* or *formal system*, being a language just the list of the alphabet and the recursively defined strings on it.

proach), while the latter in explicating basic principles governing the concept of truth (syntactical approach).

This distinction should not be over-emphasized because the two approaches are deeply interwoven. Anyway, we are going to deal with axiomatic theories of truth and I plan to justify this choice throughout the discussion. But some reasons can already be stressed. First, axiomatic theories of truth require a very weak logical framework and they suffice for spell out truth properties in full. Moreover, we seek a theory of truth for our language, i.e. a language that includes the language in which the theory is expressed. Axiomatic theories of truth do exactly this: although they do not provide nothing but an operational meaning of the truth predicate, they do not transcend the language in which T is stated. If, as I argue, the aim of a truth theory is to investigate the notion of truth for a natural language, then the most suitable setting to do this seems to be a syntactical one. Embodying truth in a theory of syntax does justice to our presentation of truth as a linguistic notion, whose bearers are linguistic entities (meaningful sentences). Another important reason is the following: given axiomatizations may be varied in natural ways by dropping conditions or extending general principles. I argue that this freedom is the most important gain of dealing with axiomatic theories. Last but not least, one can compare axiomatizations as to their proof-theoretical strength, using an extensive body of well-established metamathematical techniques. Anyway, model-theoretic approaches are important to clarify specific properties of the notion of truth, to investigate models of axiomatic truth theories and to analyse their proof theory. Therefore, semantical and axiomatical approach complement one another.

Summarizing, the strategy for a logical and syntactical approach to truth is the following:

1. reduce issues concerning truth to ones concerning the logico-linguistic predicate ‘is true’;

2. analyse the truth predicate in natural languages by studying its behaviour in formalized languages;

3. fix a formalized language, the so-called base theory;

4. if the base theory is an arithmetic theory, be aware of Tarski’s theorem of undefinability of truth;

5. expand the base theory (trying to preserve the consistency) by adding the truth predicate, its axioms and rules of inference.

In what follows I shall focus on steps 3., 4. and 5.: the second section will describe the base theory most widely used in literature, PA; the third section I shall consider the inconsistent theory obtained by extending PA with a truth predicate and all the Tarski's biconditionals. This will bring us to consider other strategies of axiomatization in the fourth section.

1.2 Base theory and coding

Once we take truth to be a predicate, we need a theory to govern objects in its extension. In other words, we have to provide a formal framework (a mathematical fragment of a natural language) in which to embed a theory of truth: the base theory. Truth theories can be built on many different systems, usually a well-known mathematical theory such as PA, weak subtheories of PA or set theories etc. are used. There are many advantages in using PA as base theory: it is neither too weak nor to strong to be combined with truth axioms.

Definition 1.2.1. The first-order language $\mathcal{L}_{\mathsf{PA}}$ of Peano Arithmetic contains as logical vocabulary propositional operators (connectives) and quantifiers, an infinite stock of variables ranging over natural numbers and the identity relation $=$. The mathematical vocabulary of $\mathcal{L}_{\mathsf{PA}}$ consists of a successor function symbol s, a zero constant 0, and function symbols $+$ and $\times$ for addition and multiplication. The set ω of natural numbers is the set which contains 0 and it is closed under the successor function. We assume that for each $n \in \omega$, there is a closed term $\bar{n} \in \mathcal{L}_{\mathsf{PA}}$, a numeral, which represents it. The inductive definitions of terms and formulae are built in the usual manner. We consider the smaller-then relation $<$ as a defined predicate. The language is supposed to contain a symbol for each primitive recursive function as well. As logical axioms we take some standard formulation of classical first-order logic. The non logical axioms of PA are:

(PA1) $\neg\exists x\ [s(x) = 0]$

(PA2) $\forall x \forall y\ [s(x) = s(y) \rightarrow x = y]$

(PA3) $\forall x\ [x + 0 = x]$

(PA4) $\forall x \forall y\ [x + s(y) = s(x + y)]$

(PA5) $\forall x\ [x \times 0 = 0]$

(PA6) $\forall x \forall y\ [x \times s(y) = (x \times y) + x]$

(PA7) $\phi(0) \wedge \forall x(\phi(x) \rightarrow \phi(s(x))) \rightarrow \forall x \phi(x) \qquad$ for all $\phi(x) \in \mathcal{L}_{PA}$

The last of these axiom is an *axiom schema*: it stands for an infinite collection of axioms obtained by instantiating it with each formula of the language. This schema, which is called principle of *mathematical induction*, is known to be equivalent to the *least-number* principle:

$$\exists x \phi(x) \rightarrow \exists x(\phi(x) \wedge \forall y < x \neg\phi(y)).$$

It should be noted that by adding a new predicate to the language of PA new formulae are created and they have to be considered as instances of (PA7). We call PAT the system based on the language $\mathcal{L}_T = \mathcal{L}_{\mathsf{PA}} \cup \{T\}$. PAT is given by the axioms of PA and all instances of the induction schema with the truth predicate. For each system S extending PA and formulating in $\mathcal{L}_T$, $\mathsf{S}\restriction$ will be the system itself without any induction axioms containing the truth predicate.

The objects of our base theory are numbers. How can we speak about sentences or about linguistic objects on the whole? We have to fix some coding of the language to which T is to be added. For any language there is a way to associate a number to each linguistic expression: the method of *arithmetization* or *gödelization* of syntax. It allows us to reasoning on the expressions of some logico-mathematical language in an arithmetic theory. So, a number theory as PA has the peculiarity of being the theory we want to code and, at the same time, the theory we use to code. The starting point is a one-to-one mapping of the alphabet into natural numbers. As consequence, each word (sentence) will correspond to one and only one natural number sequence (sequence of natural number sequences). Furthermore, sequences of natural numbers can be coded by numbers. Thus, each expression of the language ϕ is expressed by natural numbers. There are different strategy to do this, but for our purposes it does not really matter which one is used.

We need, moreover, to talk about relations and operations of the syntax of $\mathcal{L}$ within a theory based on it. This can be done: by using arithmetization, syntactical and morphological concepts can be treated as collection of numbers (the codes of the object that fall under that concept). For a given language $\mathcal{L}_{\mathsf{PA}}$ we introduce the formal representation of some primitive recursive relations for its syntactical notions:

$\mathrm{Term}_{\mathcal{L}_{\mathsf{PA}}}(x)$ $(\mathrm{CT}_{\mathcal{L}_{\mathsf{PA}}}(x))$ $\Leftrightarrow$ x is a code of a (closed) term of $\mathcal{L}_{\mathsf{PA}}$;

$\mathrm{Var}_{\mathcal{L}_{\mathsf{PA}}}(x)$ $\Leftrightarrow$ x is a code of a variable of $\mathcal{L}_{\mathsf{PA}}$;

$\mathrm{Fml}_{\mathcal{L}_{\mathsf{PA}}}(x)$ $(\mathrm{AtFml}_{\mathcal{L}_{\mathsf{PA}}}(x))$ $\Leftrightarrow$ x is a code of an (atomic) formula of $\mathcal{L}$;

$\mathrm{Sent}_{\mathcal{L}_{\mathsf{PA}}}(x)$ $(\mathrm{AtSent}_{\mathcal{L}_{\mathsf{PA}}}(x))$ $\Leftrightarrow$ x is a code of an (atomic) sentence of $\mathcal{L}_{\mathsf{PA}}$.

I shall omit the subscript $\mathcal{L}_{\mathsf{PA}}$ when the context allows it.

The same holds for syntactical operations: $\mathcal{L}_{\mathsf{PA}}$ has primitive recursive representations for them.

$$\begin{array}{rlrl}
\underset{\cdot}{\neg}\ulcorner\phi\urcorner = \ulcorner\neg\phi\urcorner, & & \ulcorner\phi\urcorner\underset{\cdot}{\rightarrow}\ulcorner\psi\urcorner = \ulcorner\phi\rightarrow\psi\urcorner, & \\
\ulcorner\phi\urcorner\underset{\cdot}{\vee}\ulcorner\psi\urcorner = \ulcorner\phi\vee\psi\urcorner, & & \ulcorner\phi\urcorner\underset{\cdot}{\wedge}\ulcorner\psi\urcorner = \ulcorner\phi\wedge\psi\urcorner, & \\
\underset{\cdot}{\forall}(\ulcorner v_k\urcorner,\ulcorner\phi\urcorner) = \ulcorner\forall v_k\phi\urcorner, & & \underset{\cdot}{\exists}(\ulcorner v_k\urcorner,\ulcorner\phi\urcorner) = \ulcorner\exists v_k\phi\urcorner, & \\
\ulcorner t\urcorner\underset{\cdot}{=}\ulcorner s\urcorner = \ulcorner t = s\urcorner, & & \underset{\cdot}{R}(\ulcorner t_1\urcorner,\ldots,\ulcorner t_k\urcorner) = \ulcorner R(t_1,\ldots,t_k)\urcorner, & \\
\mathrm{num}(n) = \ulcorner n\urcorner, & & & \\
sub(\ulcorner e\urcorner,\ulcorner t\urcorner,\ulcorner v_k\urcorner) = \ulcorner e[t/v_k]\urcorner. & & &
\end{array}$$

In these definitions $t, s, t_1, \ldots, t_k$ are terms, ϕ and ψ formulae, v_k is a variable, R is a k-ary predicate symbol and $e[t/v_k]$ is the result of replacing in e each free occurrence of a variable v_k by a term t. When is it clear from the context which variable is replaced with t I write $\ulcorner\phi(\underset{\cdot}{t})\urcorner$. I write $\ulcorner\phi(\dot{x})\urcorner$ for $sub(\ulcorner\phi(v)\urcorner, \mathrm{num}(x), \ulcorner v\urcorner)$, the expression of $\mathcal{L}_{\mathsf{PA}}$ standing for the operation of replacing in ϕ the variable v with the x-th numeral. These definitions are extended in the obvious way for cases with more than one free variable. We also need a function which provides the value of a term t, $\mathrm{val}(t) = t^{\circ}$, which is a number if t is closed term. Some other notational remarks: I write $\forall t\ \phi(t)$ as an abbreviation of $\forall x\ (\mathrm{CT}_{\mathcal{L}}(x) \rightarrow \phi(x))$ and I will write $\mathrm{Sent}_T(x\underset{\cdot}{\vee}y)$, $\mathrm{Sent}_T(x\underset{\cdot}{\wedge}y)$ or $\mathrm{Sent}_T(x\underset{\cdot}{\rightarrow}y)$ instead of $\mathrm{Sent}_T(x) \wedge \mathrm{Sent}_T(y)$ since the former formulations and the latter can be proved to be equivalent in PA.

1.3 Tarski's theorem and the Liar

Given a base theory, the first move in building a truth theory might be to find a formula T that explicitly defines the truth predicate. According to Tarski, a truth definition should meet an adequacy condition: saying 'the sentence ϕ is true' should be the same of saying 'ϕ'. Although it seems to be a very intuitive and acceptable condition — it mirrors the use we make of the truth predicate in our natural language — the claim that there is a formula $T(x)$ satisfying all of them cannot be fulfilled. This limitative results is the content of the Tarski's undefinability theorem. This theorem follows as corollary from the *diagonal lemma* or *fixed point theorem*, which establishes the existence of self-referential sentences in a formal system in which a Gödelnumbering is available. Such systems have to contain a portion of arithmetic, at least the system of *minimal arithmetic* Q, in which every recursive function is representable [6].

Theorem 1.3.1 (Diagonal lemma — Gödel, 1931)**.** Let S be a formal system containing Q. For every formula $\phi(x) \in \mathcal{L}_{\mathsf{S}}$, there is a sentence $\beta \in \mathcal{L}_{\mathsf{S}}$ such

[6] For an explanation and details see Boolos and Jeffrey [4], chapters 16 and 17.

that:

$$\mathsf{S} \vdash \beta \leftrightarrow \phi(\ulcorner\beta\urcorner/x),$$

where $\phi(\ulcorner\beta\urcorner/x)$ is the result of substituting in ϕ for the variable x the code numeral for β.

The diagonal lemma says that the diagonal function is expressible in any system containing a small portion of arithmetic; namely, in a very intuitive and common way of saying: if S is enough powerful then sentences of S can coherently talk about themselves. More formally, a sentence β is S-provably equivalent to another sentence that attributes to β a property ϕ. That is why, it is usually said that 'β says about itself that it has the property expressed by ϕ'. The sentence β can also be viewed as a fixed point of the formula $\phi(x)$, or better a fixed point of the operation assigning to each formula ψ the sentence $\phi(\ulcorner\psi\urcorner)$.

The theorem 1.3.1 yields the undefinability theorem:

Theorem 1.3.2 (Undefinability Theorem — Tarski, 1936)**.** No consistent extension S of Q proves

$$\phi \leftrightarrow T(\ulcorner\phi\urcorner)$$

for any sentence $\phi \in \mathcal{L}_{\mathsf{S}}$.

Proof. In order to obtain a contradiction, we assume that there is a consistent formal system S extending Q which proves all Tarski-biconditionals, then in particular it proves

$$\lambda \leftrightarrow T(\ulcorner\lambda\urcorner),$$

where λ is a liar-like sentence, that is a sentence that says of itself that it is false. We use the extended diagonal lemma to produce the liar sentence such that:

$$\mathsf{S} \vdash \lambda \leftrightarrow \neg T(\ulcorner\lambda\urcorner).$$

These equivalences together yield a contradiction in S:

$$\mathsf{S} \vdash T(\ulcorner\lambda\urcorner) \leftrightarrow \neg T(\ulcorner\lambda\urcorner).$$

□

The notion of truth, as it is formalized by Tarski-biconditionals, cannot be internalized within an arithmetical theory: such systems cannot define their own truth. In proving this, the causal role of the Liar comes out. Contradiction is engendered by formalizing the liar sentence through diagonalization: λ 'says of itself' that is false. This conflicts with the Tarskian adequacy condition for truth. Therefore the only viable way is to move to step 5. of the initial list: expanding the base theory (trying to preserve the consistency) by adding the truth predicate as a primitive, non-defined one. The interpretation of this predicate is fixed by providing axioms and rules of inference.

In doing this we have to keep a close eye on the inconsistency problem, as we shall see immediately.

The most immediate and natural choice seems to take T-biconditionals as axioms. Let us consider the system PAT based on the language $\mathcal{L}_T = \mathcal{L}_{\mathsf{PA}} \cup \{T\}$ without any truth axiom and expand it with an axiom schema:

$$\phi \leftrightarrow T(\ulcorner \phi \urcorner) \qquad \text{for any sentence } \phi(x) \in \mathcal{L}_T.$$

That is to say, given the impossibility of finding an $\mathcal{L}_{\mathsf{PA}}$-formula such that all the Tarski-biconditionals for the language $\mathcal{L}_T$ are proved, a new predicate is added and they are simply taken as axioms. By doing so we have built an axiomatic theory of truth that we call NT, the *naive theory of truth*, following Horsten[7]. This theory is trivially inconsistent for the theorem 1.3.2, nevertheless I judge it interesting for heuristic purposes, as far as this theory to some extent formalize the liar paradox and so, it allows an analysis of the paradox itself.

In order to carry out this analysis, our starting point is the liar paradox in natural languages; let us take it in its most famous version: "I am lying now" or "This sentence is false"[8]. It is worth noting that the formalized version of the paradox is a far greater threat for a formal system than it is its intuitive version for natural language. Tarski's diagnosis in [46] about the Liar is that natural language is infected by contradiction. But there is no agreement about it: can natural language be regarded as inconsistent? An interesting stance, that I can just mention, is due to Burge. He argued that when we blame natural languages for contradictions actually we are referring to something that is no more natural but already 'theory-laden':

> Natural languages *per se* do not postulate or assert anything. What engenders paradox is a certain naive theory or conception of the natural concept of truth[9].

Hence, in his opinion, paradoxes do not affect the language itself but rather our very intuitive theories about the language. As far as formal systems are concerned, there is no way out: if the Liar can be carried out in a formal system, the latter becomes inconsistent. At any rate, for our discussion it does not really matter whether conditions that let the paradox arise are considered as related either to the language or to part of an intuitive theory about language: they should be viewed only as a bridge for the analysis of the paradox in formal systems.

Following the notorious Tarskian diagnosis[10], it is possible to isolate the roots of the paradox, i.e conditions without which it would not be generated:

[7] Cfr. Horsten [31], p. 55.

[8] These sentences are roughly formulated, anyway it is not a problem to write down a non-contingent liar sentence.

[9] Cfr. Burge [5], p. 179.

[10] Cfr. Tarski [46], p.165.

1. Our ordinary language has a quotational mechanism (which gives a name to each sentence). Moreover, it admits self-referential sentences such as “This sentence is written in English” and, furthermore self-referential sentences which involve the notion of truth such as “This sentence is true”.

2. We implicitly accept Convention T, namely the fact that the truth of a sentence is reduced to the sentence itself and *vice versa*.

3. We accept ordinary reasoning, especially bivalence, i.e. the claim that a sentence must be either true or false.

These implicit conditions seems very uncontroversial and, at the same time, impossible to avoid. Natural languages, even though rich and interesting, do not admit the needed handling for an analysis of the paradox. To address this issue by using logic as a tool is a fascinating strategy since it gives us the chance of turning these intuitive features of natural language in formally specified conditions of a formal system. In this way, the roots of the paradox can be escape routes as well: in a formal setting one can remove or modify any of the conditions in order to avoid the paradox and to ward off the inconsistency of the theory.

Now we can turn previous conditions in formal terms. Let S be a formal system on the language $\mathcal{L}_T$, with the predicate $T(x)$ expressing that x is true, the contradiction via liar paradox is a result of the following combination of features:

1. **Syntax**:

   (i) *Naming*: each sentence ϕ of $\mathcal{L}$ has name in the language, the closed term $\ulcorner\phi\urcorner$ of $\mathcal{L}$.

   (ii) *Self-reference*: via diagonalization, for each formula $\phi(x)$ we can find a sentence β which is provable equivalent in S to $\phi(\ulcorner\beta\urcorner)$.

2. **Basic principles**: Tarski biconditionals are accepted for each sentence of $\mathcal{L}_T$.

3. **Logic**: The underlying logic is a standard formulation of classical first-order predicate calculus[11].

Of course, these properties are desirable for a theory of truth that aims to be as close as possible to natural language. But at the same time, they cannot be all simultaneously satisfied on pain of inconsistency. NT meets all stated condition, hence it is inconsistent. At this point the question is: how can we do better than NT?

[11] Actually since the law of excluded middle is not used in the derivation of the paradox, the argument is already derivable in *minimal logic*.

1.4 How to face the Liar

Following Feferman[12], I said that the roots of the paradox once isolated in a formal system may also represent the possible ways out: by dropping somehow one of them one can ensure the consistency of the theory. Corresponding to 1.–3. there are three kinds of restriction:

1*. Restriction of syntax.

2*. Restriction of basic principles.

3*. Restriction of logic.

The choice of a strategy does not uniquely determine a theory, because restrictions can be designed in different ways. And indeed it was so, historically many different and consistent truth theories have been developed. The *fil rouge* I shall follow in presenting theories is their connection with the strategies put in place to avoid the paradox. This reflects a definite belief about the philosophical value of truth theories: they are possible solutions for semantical paradoxes. This is why we are interested in the way in which they 'solve', actually 'go around', the Liar. We shall follow the three points above to orient ourselves in the constellation of axiomatic theories of truth; however, I am going to provide a very brief account, focusing on theories that will be useful throughout our discussion[13].

1.4.1 Restriction of syntax

Let us have a closer look to the first restriction. The process of naming in formalized language is a tool that simulates the device of quotation in natural language: if truth is seen as a property of sentences there should be linguistic objects representing them. In more formal terms, truth is a monadic predicate T which applies to closed terms, namely numerals of gödelian of sentences. What about self-reference? Is it dispensable at least for formulae containing the truth predicate? The strategy might be to impose restrictions to the language, particularly on the formation of formulae by forcing a sort of stratification. This is the Tarskian strategy, the most immediate and simple solution to the Liar. In this way, the language about which we talk (object-language) and the language we use to talk about the former (metalanguage) are distinguished. The object-language does not contain the truth predicate, it just contains truth-free sentences and a truth definition for it has to be given in a metalanguage which is essentially stronger than it. Then, we can define the truth predicate for the metalanguage in a meta-metalanguage and so on. This process can be iterated and a hierarchy of languages can be built:

[12] Cfr. Feferman [14], p. 81

[13] For an organic presentation see Halbach [28] or Horsten [31].

a truth predicate for a language $\mathcal{L}_0$ is available only in a second language $\mathcal{L}_1$, whose truth predicate in turn must belong to a language $\mathcal{L}_2$ etc. Hence $T\ulcorner\phi\urcorner$ is a sentence of a language $\mathcal{L}_n$ if and only if ϕ is a sentence of $\mathcal{L}_{n-1}$, in this way a truth predicate is only applied to formulae containing variables that range over arithmetic (i.e. truth free) formulae or over formulae of the previous language in the hierarchy. So, liar-like sentences are no longer well-formed and, accordingly, they cannot be substituted in the T-schema. This prevents contradiction from arising.

However, as it has been repeatedly emphasized, this strategy seems to be excessively restrictive for at least three reasons:

- It rules out unproblematic sentences as well; sentences that should be admissible, or at least expressible in a language.
- Regarding problematic sentence, to require that they are not sentences at all would be like uprooting a plant that should be just pruned. In fact, sentences as the liar one become critical only when they are submitted to the disquotational device.
- In natural languages there is no sign of this stratification. Since the pretended unnaturalness of typing is open to discussion, the last statement has to be justified. It was already said and it will be repeated that since we are working in an artificial framework to advocate issues of naturalness would be needlessly restrictive. Nevertheless, I argue that, as far as this kind of typing (with syntactic restriction) is concerned, we move away overmuch from the ordinary language. The language $\mathcal{L}_T$, though artificial, keeps the familiar shape of other languages: there are words, sentences and truth is a property of *whatever* (codes of) sentences. Stratification of the language would drastically diminish the chance of apply formal theory to informal language. This is not to say that the aim is to reconstruct natural language in a formal theory, but to preserve that loose syntactic resemblance could help once a philosophical approach to axiomatic theories of truth is adopted. To this aim, an intended interpretation of a truth theory should be retained: terms stand for whatever sentence and the truth predicate expresses a property of these sentences. Therefore, I argue that we should use an *untyped truth predicate*; note that this does not rule out *typed truth theories.* Unlike syntactic restriction, there are other kinds of typing that seem less problematic: the truth predicate has an 'operational definition' given by providing axioms, so principles governing its action should be formulated without restraint and without the ghost of naturalness.

Hence, a better strategy seems not to limit the expressive power, understood strictly, of the theory, but to restrict the application of Tarski-biconditionals.

This brings us to the second way: the restriction of basic principles. As we shall see, the distinction between 1* and 2* is fine but nevertheless significant: on the one hand, we have a restriction in the inductive definition clauses for formulae, on the other hand, language is left unchanged but the axioms are weakened. Therefore, the Tarskian hierarchical approach moves from the language to the basic principles, in other words one allows the presence of sentences with iteration of truth predicate but axioms concern just truth free sentences. However, the results is the same: a *typed* theory of truth is built, that is a theory that does not allow self-reference. So, there are two ways in which one can obtain a typed system: with a syntactic typing (as in the way 1*) or by formulating its axioms in a way that renders it a typed system (2*). I argue that this distinction has to be taken in account talking about the naturalness of typing.

At any rate, typing is not at all the only solution for the liar: still remaining within the strategy 2* we shall see there are other ways to bound the axioms in such a way that the resulting theory is *type-free*. As Halbach[14] noted, there is a sufficient condition for a theory's being type-free: it has to *prove the truth* of at least a sentence containing the truth predicate. Type-free theories are also called theories of *self-referential* truth.

1.4.2 Restriction of basic principle

Once a restriction of principles is chosen, the obvious principle needing restriction is the T-schema as adding all Tarski-biconditionals to a base theory that allows the diagonalization leads to contradiction. At the same time, in restricting basic principles there is the attempt of limiting as little as possible the expressive power of the theory and preserving interesting features of T such as compositionality or iteration. I now sketch possible ways for bounding or weakening T-schema and the resulting theories[15]:

2*a. To restrict the disquotational schema to $\mathcal{L}_{\mathsf{PA}}$-sentences or $\mathcal{L}_{\mathsf{PA}}$-formulae [Theories TB and UTB].

- To add compositionality clauses for T. [Theory CT]
- To allow iteration of the truth predicate by resuming the Tarskian idea of hierarchical theories of truth [Theories $\{\mathsf{RT}_\alpha\}$].

[14] Cfr. Halbach [28], p. 145.

[15] Of course it is an *a posteriori* reconstruction based on the most widely used theories of truth. There are many different ways to pursue restriction aimed at avoiding paradoxes. Another possibility I would like to mention is due to Leitgeb, his answer for the problem of restricting biconditional is based on the semantic notion of dependence: sentences with truth predicate that can be inserted plausibly and consistently into the T-schema are those sentences which depend directly or indirectly on non-semantic states of affairs (only), in a sense formally precised. I do not include his theory in this survey as far as it is a semantic theory whereas we are dealing with syntactical ones.

2*b. To restrict the disquotational schema to special classes of $\mathcal{L}_T$-sentences [Theories PTB and PUTB].

2*c. To consider instead of T-schema the corresponding (and weaker) rules of inference [Theory FS].

Typed disquotational truth

An axiomatization of truth based on T-schema is desirable and Tarski himself provided an axiomatization of this kind. A possible route to avoid inconsistency is to restrict the disquotational schema to sentences of $\mathcal{L}_{\mathsf{PA}}$ without the truth predicate. In this way we obtain a typed system in which no sentence of the form $T\ulcorner\phi\urcorner$ with $\phi \in \mathcal{L}_T$ can be proved, this system is called TB for *Tarski-biconditionals.*

Definition 1.4.1. The theory TB comprises all axioms of PAT (PA formulated in $\mathcal{L}_T$) and, moreover, all sentences of the form $T\ulcorner\phi\urcorner \leftrightarrow \phi$ where ϕ is a sentence of the language of $\mathcal{L}_{\mathsf{PA}}$.

In this theory, the formulae ϕ are sentences, but this request can be released and the theory strengthened. If the formulae ϕ are allowed to contain free variables, the theory UTB— acronym for *uniform Tarski-biconditionals* — is obtained.

Definition 1.4.2. The theory UTB comprises all axioms of PAT and all sentences of the form

$$\forall t_1, \ldots, t_n (T\ulcorner\phi(\underset{\cdot}{t}_1, \ldots, \underset{\cdot}{t}_n)\urcorner \leftrightarrow \phi({t_1}^{\circ}, \ldots, {t_n}^{\circ}))$$

where $\phi(x_1, \ldots, x_n)$ is a formula of the language of $\mathcal{L}_{\mathsf{PA}}$ with exactly $x_1, \ldots, x_n$ free.

By the conventions, t° stands for the value of the term t and $\ulcorner\phi(\underset{\cdot}{t})\urcorner$ stands for $\ulcorner\phi(\underset{\cdot}{t}/\ulcorner x\urcorner)\urcorner$, i.e. the result of substituting all free occurrence of the variable x with t in $\phi(x)$. Intuitively, biconditional displayed in the definition says that $\phi(t_1, \ldots, t_n)$ is true if the relation denoted by ϕ holds between the terms denotations. Theories TB$\restriction$ and UTB$\restriction$ are, respectively, the theories TB and UTB with the induction schema restricted to the language $\mathcal{L}_{\mathsf{PA}}$.

It can be showed that in TB the liar paradox cannot be derived: the liar sentence λ belongs to $\mathcal{L}_T$ and not to $\mathcal{L}_{\mathsf{PA}}$, so it is not a permissible substitution in the T-schema. Although this is not enough to state its consistency — it is a necessary but not sufficient condition — we can find a model based on natural numbers for TB, hence, by the soundness theorem, TB is consistent. So TB has some interesting features, among which we can mention the conservativity over PA, nevertheless it is to some extent deductively weak. As said before, that with the help of truth predicate one can express infinite conjunctions and disjunctions, but it can be shown that UTB, and therefore TB, $\mathsf{TB}\restriction$ and $\mathsf{UTB}\restriction$, cannot prove any infinite generalization. For these reasons, one would aim to do better.

Typed compositional truth

Disquotational theories have the advantage of being very natural and not arbitrary axiomatizations; in spite of this they provide no full validation to another intuitive feature of the notion of truth: its *compositional nature.* For example, an intuitively plausible logical principle concerning truth might be the distributivity of truth over a logical connective. TB proves all instances of these principles, but it cannot collect them into a general theorem valid for all the sentences of $\mathcal{L}_{\mathsf{PA}}$.

These desirable principles are contained in the 'inductive clauses' employed by Tarski to define recursively the notion of truth of a formula in a model. By turning them into axioms we obtain a theory of truth which proves the sought generalizations and which is still natural: the theory CT, for *compositional truth.*

Definition 1.4.3. The theory CT comprises all axioms of PAT and the following axioms:

(CT1) $\forall s \forall t\ [T(\ulcorner s = t \urcorner) \leftrightarrow s^{\circ} = t^{\circ}]$

(CT2) $\forall x\ [\mathrm{Sent}_{\mathsf{PA}}(x) \rightarrow (T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T x)]$

(CT3) $\forall x \forall y\ [\mathrm{Sent}_{\mathsf{PA}}(x \underset{\cdot}{\wedge} y) \rightarrow (T(x \underset{\cdot}{\wedge} y) \leftrightarrow T(x) \wedge T(y))]$

(CT4) $\forall x \forall y\ [\mathrm{Sent}_{\mathsf{PA}}(x \underset{\cdot}{\vee} y) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow T(x) \vee T(y))]$

(CT5) $\forall x \forall y\ [\mathrm{Sent}_{\mathsf{PA}}(x \underset{\cdot}{\rightarrow} y) \rightarrow (T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow (T(x) \rightarrow T(y)))]$

(CT6) $\forall v \forall x\ [\mathrm{Sent}_{\mathsf{PA}}(\underset{\cdot}{\forall} v x) \rightarrow (T(\underset{\cdot}{\forall} v x) \leftrightarrow \forall t T(x[t/v]))]$

(CT7) $\forall v \forall x\ [\mathrm{Sent}_{\mathsf{PA}}(\underset{\cdot}{\exists} v x) \rightarrow (T(\underset{\cdot}{\exists} v x) \leftrightarrow \exists t T(x[t/v]))]$

Note that this system is typed as the quantifiers range over sentences of $\mathcal{L}_{\mathsf{PA}}$, which are truth-free sentences. If $\mathcal{L}_T$-sentences were included in the axioms as well, the unrestricted Tarski-biconditionals would be derivable and the system would be inconsistent. The system is compositional in the sense that the truth of a sentence depends on the values of the constituents of that sentence.

Hierarchical compositional truth

Both disquotational and compositional theories do not prove sentences which contain a truth iteration, such as

$$T^{\ulcorner} T^{\ulcorner} 0 = 0^{\urcorner \urcorner},$$

as they are typed system based on the language $\mathcal{L}$. Now, we can consider PAT as base theory and add a new truth predicate whose domain ranges over sentences of $\mathcal{L}_T$. This predicate, in order to preserve the consistency, has to be a new predicate T_1. The old predicate T is treated merely as a predicate symbol of the base language $\mathcal{L}_T$[16]. The process can be iterated by adding at each step a new predicate up to the point that the truth predicates can be indexed by natural numbers $T_0, T_1, T_2, \ldots$. At this point, a further theory of truth can be formed with another truth predicate T_ω, and so on for further ordinal numbers. This hierarchy can go up beyond transfinite levels, the problem is that such a theory cannot be formalized because of problems concerning the coding. A possible solution is to build a hierarchy along an initial segment of natural numbers up to a halting point, the ordinal Γ_0[17]. The language extended with all the predicate $\{T_i\}_{i<\Gamma_0}$ can be coded and the corresponding theory of truth formalized. I indicate the ordinals with Greek letters and, if α is an ordinal, then $\bar{\alpha}$ is the numeral of its code.

Definition 1.4.4. For an ordinal $\gamma \leq \Gamma_0$ the language $\mathcal{L}_{<\gamma}$ is $\mathcal{L}_{PA}$ expanded by all truth predicates T_β for all $\beta < \gamma$; adding T_γ as well, the language $\mathcal{L}_\gamma$ is obtained.

For any ordinal γ smaller or equal to Γ_0 a theory $\mathsf{RT}\ _{<\gamma}$ is defined. The label stands for theory of *ramified truth up to* γ.

Definition 1.4.5. For $\alpha \leq \Gamma_0$ the theory $\mathsf{RT}_{<\alpha}$ is given by all the axioms of PA, induction axioms for $\mathcal{L}_{<\alpha}$ and, for all $\gamma < \beta < \alpha$:

(RT1) $\forall s \forall t\ [T_\beta \ulcorner s = t \urcorner \leftrightarrow s^{\circ} = t^{\circ}]$

(RT2) $\forall x\ [\mathrm{Sent}_{<\beta}(x) \rightarrow (T_\beta(\underset{\cdot}{\neg} x) \leftrightarrow \neg T_\beta x)]$

(RT3) $\forall x \forall y\ [\mathrm{Sent}_{<\beta}(x \underset{\cdot}{\vee} y) \rightarrow (T_\beta(x \underset{\cdot}{\vee} y) \leftrightarrow T_\beta(x) \vee T_\beta(y))]$

(RT4) $\forall x \forall y\ [\mathrm{Sent}_{<\beta}(x \underset{\cdot}{\rightarrow} y) \rightarrow (T_\beta(x \underset{\cdot}{\rightarrow} y) \leftrightarrow (T_\beta(x) \rightarrow T_\beta(y)))]$

[16] Note that, generally speaking, if a new predicate P is added to the language of the base theory in TB it is enough to add new instances of the disquotational schema, while in CT a new axiom analogous to the axiom (CT1) must be added as well:

$$\forall t (T(\underset{\cdot}{P}(t)) \leftrightarrow P(t^{\circ})).$$

In this case, the theory CT is based on the language $\mathcal{L}_T \cup \{T_1\}$ and the predicate 'added' to the base theory is T, hence, the new axiom must be:

$$\forall t (T_1(\underset{\cdot}{T}(t)) \leftrightarrow T(t^{\circ})).$$

[17] Γ_0, called Feferman-Schütte ordinal, is a large countable ordinal, for its precise definition see Feferman [13]. It has an important role since is the smallest ordinal that cannot be defined from smaller ordinals by using 'predicative' means. For this reason, it is commonly considered the limit of predicativity.

(RT5) $\forall v \forall x\ [\mathrm{Sent}_{<\beta}(\underset{.}{\forall} v x) \rightarrow (T_\beta(\underset{.}{\forall} v x) \leftrightarrow \forall t T_\beta(x[t/v]))]$

(RT6) $\forall t\ [\mathrm{Sent}_{<\gamma}(t^\circ) \rightarrow (T_\beta(\underset{.}{T}_\gamma t) \leftrightarrow T_\gamma t^\circ)]$

(RT7) $\forall t \forall \delta \prec \bar{\beta}\ [\mathrm{Sent}_{<\delta}(t^\circ) \rightarrow (T_\beta(\underset{.}{T}_\delta t) \leftrightarrow T_\beta t^\circ)]$

The first six axioms are generalization of the axiom of CT: each truth predicate T_β with $\beta < \gamma$ satisfies the Tarski's inductive clauses. The axiom (RT7) is a generalization of the axiom (CT1) that governs the truth of atomic sentences[18]. The last axiom is specific for a theory with more than one truth predicate and expresses their cumulativity: the truth predicates T_α and T_β agree on sentences of the language $\mathcal{L}_\alpha$ if $\alpha < \beta$, therefore the higher truth predicates include all the lower ones and the higher languages are richer than the lower ones in the hierarchy.

Type-free disquotational truth

Anyway, as I have already said, typing is not the only solution, indeed strategies 2*b. and 2*c. lead to type-free systems.

Let us came back to disquotational theories; if one drops the restriction that truth only applies to sentences without the truth predicate, then one can obtain very strong theories and prove many general principles. The problem is that the result of releasing the restriction on TB is an inconsistent theory, hence in order to obtain a type-free disquotational theory we have to find other kind of restriction on the instances of the T-schema. In other words, we need a criterion to exclude some (as little as possible) problematic instances of the T-schema from the axioms of the theory. This can be done in several ways.

The domain of $T-$biconditionals that we are looking for must be natural and as larger as possible, thus one might adopt a maximality principle adding as many disquotational sentences as is it consistently possible. Nevertheless, McGee showed that there are $2^{\aleph_0}$ many sets Γ of disquotational sentences maximally consistent with TB, i.e. such that any set $\Gamma' \supset \Gamma$ of disquotational sentences is inconsistent with TB. As consequence, we would have uncountably many different theories. Unfortunately, the same applies if one choose a criterion of maximal conservativity: there are $2^{\aleph_0}$ maximal conservative extensions of TB. There is another way: we can try to understand how to avoid the liar-like paradoxes. An essential feature of them is the presence of a *negated* occurrence of the truth predicate. It has been presented a type-free disquotational theory, which is based precisely on this insight: we can avoid the paradox by banning instances of the disquotational schema in which the truth predicate does not occur positively. A formula ϕ of $\mathcal{L}_T$ is said positive with respect to T if and only if T occurs in ϕ in the scope of an even number of negation symbols, if any.

[18] See the footnote 16 on the preceding page.

Definition 1.4.6. The system PTB is the system PAT with, as axioms, all sentences of the form $T\ulcorner\phi\urcorner \leftrightarrow \phi$ where ϕ is a T-positive sentence of $\mathcal{L}_T$.

The acronym PTB stands for *positive T-biconditionals.* Unproblematic self-referential sentences, such as the truth teller sentences $\tau := T(\ulcorner\tau\urcorner)$, are all T-positive, hence sentences like $T(\ulcorner\tau\urcorner) \leftrightarrow \tau$ are axioms of PTB. In the same way, the domain of uniform T-biconditionals can be expanded:

Definition 1.4.7. The theory PUTB comprises all axioms of PAT and all sentences of the form

$$\forall t_1, \ldots, t_n(T\ulcorner\phi(\underset{\cdot}{t}_1, \ldots, \underset{\cdot}{t}_n)\urcorner \leftrightarrow \phi(t_1{}^{\circ}, \ldots, t_n{}^{\circ}))$$

where $\phi(x_1, \ldots, x_n)$ is a formula of the language of $\mathcal{L}_T$ with exactly $x_1, \ldots, x_n$ free and in which T occurs only positively.

Type-free compositional truth

It is well-known that an axiom with an implicative form such as $\vdash A \to B$ is always stronger than the corresponding rule $\vdash A \Rightarrow \vdash B$. For this reason one may wonder whether by substituting T-schema with corresponding rules he gets a consistent theory. More explicitly, the two directions of the schema $\phi \to T(\ulcorner\phi\urcorner)$ and $T(\ulcorner\phi\urcorner) \to \phi$ can be replaced with the matching rules:

$$\frac{\phi}{T(\ulcorner\phi\urcorner)}\ \mathsf{Nec}$$

$$\frac{T(\ulcorner\phi\urcorner)}{\phi}\ \mathsf{Conec}$$

The label Nec stands for *necessitation rule*, for similarity with its modal analogue; Conec stands for *conecessitation rule.*

In order to avoid the inconsistency is it enough to replace one direction of the disquotational schema with the matching rule? The answer is negative, the liar paradox can be strengthened as follow:

Theorem 1.4.1 (Montague's theorem)**.** Any system extending PA closed under the rule Nec for all sentences of $\mathcal{L}_T$ and under the schema $T(\ulcorner\phi\urcorner) \to \phi$ for all sentences ϕ of $\mathcal{L}_T$ is inconsistent.

Theorem 1.4.2 (Dual of Montague's theorem)**.** Any system that contains PA and the schema $\phi \to T(\ulcorner\phi\urcorner)$ for all sentences ϕ of $\mathcal{L}_T$ is inconsistent with the rule Conec.

However, there is a good news: Nec and Conec can be consistently combined with one another, and also with many further axioms. If a truth theory is closed under these two rules, it is called *symmetric* and it can be proved that the liar sentence is neither provable nor refutable in it.

Exploiting this chance a type-free generalization of CT is built, by dropping the restriction on $\mathcal{L}$-sentences and by postulating the commutativity between truth and logical operator even for sentences which contain the truth predicate and not just for arithmetical sentences. The resulting system is called FS. The label stands for *Friedman–Sheard*, who fist presented this theory in [21]. The following formulation is due to Halbach [28].

Definition 1.4.8. The system FS is given by all axioms of PAT, the following axioms:

(FS1) $\forall s \forall t\ [T\ulcorner s = t\urcorner \leftrightarrow s^{\circ} = t^{\circ}]$

(FS2) $\forall x\ [\mathrm{Sent}_T(x) \rightarrow (T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T x)]$

(FS3) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\wedge} y) \rightarrow (T(x \underset{\cdot}{\wedge} y) \leftrightarrow T(x) \wedge T(y))]$

(FS4) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\vee} y) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow T(x) \vee T(y))]$

(FS5) $\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\forall} v x) \rightarrow (T(\underset{\cdot}{\forall} v x) \leftrightarrow \forall t T(x[t/v]))]$

(FS6) $\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\exists} v x) \rightarrow (T(\underset{\cdot}{\exists} v x) \leftrightarrow \exists t T(x[t/v]))]$

and the following inference rules: for each sentence ϕ of $\mathcal{L}_T$:

$$\frac{\phi}{T(\ulcorner \phi \urcorner)}\ \mathsf{Nec}$$

$$\frac{T(\ulcorner \phi \urcorner)}{\phi}\ \mathsf{Conec}$$

FS has of course desirable properties, but the most relevant result concerning FS is the fact that it is ω-inconsistent.

Definition 1.4.9. A theory S is ω-inconsistent if and only if there is a formula $\phi(x)$ such that $\mathcal{S} \vdash \phi(\bar{n})$ for every $n \in \omega$ and $\mathcal{S} \vdash \neg \forall x \phi(x)$.

The inconsistency of FS is an application of a negative result due to McGee:

Theorem 1.4.3 (McGee's theorem)**.** Any theory S containing all the axioms of PA that is closed under NEC and proves the following sentences is ω-inconsistent:

(i) $\forall x(\mathrm{Sent}_T(x) \rightarrow (T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T x))$;

(ii) $\forall x \forall y(\mathrm{Sent}_T(x \underset{\cdot}{\vee} y) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow T(x) \vee T(y))$;

(iii) $\forall v \forall x(\mathrm{Sent}_T(\underset{\cdot}{\forall} vx) \rightarrow (T(\underset{\cdot}{\forall} vx) \leftrightarrow \forall t T(x[t/v])))$.

I would like just mention the existence of an interesting debate about how dire is for a truth theory to be ω-inconsistent, there is no agreement about this. At any rate, FS turns out to have an interesting semantic, especially if subsystems of FS with a limited number of application are considered (FS $_n$)[19].

1.4.3 Restriction of logic

In order to avoid paradoxes another way out can be taken. This is based on the condition 3., namely the fact that *ordinary reasoning* is requested for deriving the paradox. A possible solution is to change somehow the underlying logic by revising logical axioms and rules. Classical logic is also characterized by the fact that every property is consistent and complete with respect to the domain. Formally, given an universe of discourse M and a property X on M, let $\bar{X}$ be the complement of X with respect of the domain. We have:

Consistency of X: $X \cap \bar{X} = \emptyset$, a set and its complement are mutually exclusive.

Completeness of X: $X \cup \bar{X} = M$, a set and its complement exhaust the domain.

These conditions for the predicate 'is true' are expressed by the so-called Consistency and Completeness axioms:

Cons:	$\forall x[\mathrm{Sent}_T(x) \rightarrow \neg(Tx \wedge T\underset{\cdot}{\neg}x)]$	No sentence is both true and false.
Comp:	$\forall x[\mathrm{Sent}_T(x) \rightarrow (Tx \vee T\underset{\cdot}{\neg}x)]$	Every sentence is either true or false.

In can be proved that Cons is incompatible with the schema $\phi \rightarrow T\ulcorner\phi\urcorner$ for all sentences of $\mathcal{L}_{\mathsf{PA}}$ and, dually, Comp cannot be added to a system which proves $T\ulcorner\phi\urcorner \rightarrow \phi$. The logic of truth must be changed and, typically, there are two strategies:

3*a. Cons is maintained and Comp dropped : T becomes a partial predicate, *truth-value gaps* are allowed, that is there are sentences neither truth nor false.

3*b. Comp is maintained and Cons dropped: *truth-value gluts* are admitted, that is there are sentences both truth and false.

[19] See Halbach [28] for further details.

Partial truth

The first suggestion in this direction, due to Kripke, is to employ three-value logic[20]. Kripke's theory is a semantic one: he provided a partial model for the language $\mathcal{L}_T$, by extending a standard model of $\mathcal{L}_{\mathsf{PA}}$ with a suitable interpretation for a type-free truth predicate. T is partially defined, that is its interpretation is given by two sets: the extension and the antiextension of the truth predicate, such that a sentence is true (is false) if and only if its code is in the extension (in the antiextension) of the truth predicate.

Since I shall comprehensively explain Kripke's construction in section 3.1.4, now I just want to tell how the Liar is avoided. In classical models the antiextension of a property coincides with the complement of the extension with respect to the domain — in our case, the set of all sentences. So, in Tarskian models for $\mathcal{L}_T$, a given sentence either holds or its negation holds. In partial models this does not happen: the union of extension and antiextension does not exhaust the domain (the set of sentences of $\mathcal{L}_T$), accordingly there are sentences that do not belong to any of the two sets, i.e. sentence neither true nor false. Kripke's model is built in order to ensure that the fate of paradoxical sentences is to fall in this gap[21]. In this way inconsistencies are avoided.

Kripke did not provide an axiomatization of his theory, Feferman in [16] gave an axiomatic formalization of it which has been later called KF, for *Kripke-Feferman.* Usually in KF formulations a further primitive predicate for falsity is used, defined by the sentence:

$$\forall x(Fx \leftrightarrow T\underset{\cdot}{\neg}x),$$

of course x ranges over numerals of gödelian of sentences. So, falsity for a sentence is defined as the truth of its negation.

Definition 1.4.10. The system KF is given by all axioms of PAT and the following axioms:

(KF1) $\forall s \forall t\ [T\ulcorner s \underset{\cdot}{=} t\urcorner \leftrightarrow s^{\circ} = t^{\circ}] \wedge \forall s \forall t\ [F\ulcorner s \underset{\cdot}{=} t\urcorner \leftrightarrow s^{\circ} \neq t^{\circ}]$

(KF2) $\forall t\ [T(\underset{\cdot}{T}t) \leftrightarrow Tt^{\circ}] \wedge \forall t\ [F(\underset{\cdot}{T}t) \leftrightarrow (T\neg t^{\circ})]$

(KF3) $\forall x\ [\mathrm{Sent}_T(x) \to (T(\underset{\cdot}{\neg}\underset{\cdot}{\neg}x) \leftrightarrow Tx)]$

(KF4) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\vee} y) \to (T(x \underset{\cdot}{\vee} y) \leftrightarrow T(x) \vee T(y))]$

(KF5) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\vee} y) \to (F(x \underset{\cdot}{\vee} y) \leftrightarrow F(x) \wedge F(y))]$

(KF6) $\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\forall} v x) \to (T(\underset{\cdot}{\forall} v x) \leftrightarrow \forall t T(x[t/v]))]$

[20] See Kripke [32].

[21] There is also an alternative solution: the overlapping of the two sets.

(KF7) $\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\forall} vx) \to (F(\underset{\cdot}{\forall} vx) \leftrightarrow \exists t F(x[t/v])))]$

It is worth noting that axiom (KF2) is incompatible with the second axiom of FS:

$$\forall x(\mathrm{Sent}_T(x) \to (T(\underset{\cdot}{\neg} x) \leftrightarrow \neg Tx)),$$

a formula which states the equivalence between the claim that a sentence is not true and the claim that the sentence's negation is true. This means that in systems like KF being not true or being false are different properties.

It can be also proved that (FS2) is logically equivalent to the conjunction of Cons and Comp taken in the following versions:

$$\begin{aligned} \mathsf{Cons}: &\quad \forall x(\mathrm{Sent}_T(x) \to (T\underset{\cdot}{\neg} x \to \neg Tx)) \\ \mathsf{Comp}: &\quad \forall x(\mathrm{Sent}_T(x) \to (\neg Tx \to T\underset{\cdot}{\neg} x)) \end{aligned}$$

So adding both of them to KF leads to contradiction. Nevertheless either of them can be added preserving consistency and $\mathcal{L}_{\mathsf{PA}}$-conservativity[22]. Moreover, resulting systems KF + Cons and KF + Comp are equivalent for their arithmetical content.

Dialethic truth

Excluding gaps and admitting only gluts leads to a conceptions which is usually called *dialethic* truth. In these theories there are sentences, typically the Liar, that are both true and false against Cons. I shall not deal with this kind of approach, but see Field [19].

1.5 Adequacy criteria

In short, in setting up a theory of truth one has to avoid liar-like contradictions and, at the meantime, to do justice to the several facets typical of the notion of truth as it is used by speakers, features which are partially embodied by condition 1.–3. . But not all of them can be satisfied at the same time, so we are led to loosen in some way them. The problem we find ourselves to face is: how far could we go from the original requirements?

I am going to deal again with the issue of naturalness, but first let us provide further materials for our discussion by seeing other criteria to test proposed solutions, as far as consistency is just a minimal criterion for theories of truth. Leitgeb, in a paper meaningfully entitled *What theories of truth should be like (but cannot be)*[23], isolates eight adequacy criteria that truth theories should meet but at the same time they cannot fully meet as otherwise they would be inconsistent.

[22] See Cantini [7].

[23] Leitgeb [33]

(a) Truth should be expressed by a predicate (and a theory of syntax should be available).
I have already justified this kind of approach and, actually, all standard theories of truth meet this criterion.

(b) If a theory of truth is added to mathematical or empirical theories, it should be possible to prove the latter true.
Suppose a theory of truth T is built by adding truth predicate and truth axioms to a base theory B. Theorems of B are of course *proved* in the combination of the two theories, but this is not enough. Since a predicate for truth is available, doubtless it would be strange if theorems of B could not be *proved true.*

(c) The truth predicate should not be subject to any type restrictions.
By saying this Leitgeb understands that kind of restriction we have shown in route 1*., that is a type-limitation for the formulae allowed to fall in the scope of T.

(d) T-biconditionals should be derivable unrestrictedly.
As Tarski suggested, checking whether all T-biconditionals for the language of the truth theory are derivable in the theory itself is a secure criterion to test how far we are from a 'good' truth definition.

(e) Truth should be compositional.
Generally speaking, compositionality for truth is a phenomenon that can be described as follows: whether or not a complex sentence is true should be determined just by whether or not its constituent sentences are true and by its logical structure. This seems very natural for truth up to the point that Tarski inductively *defined* the truth of a complex sentence in terms of the truth of its logical parts (this view is embodied in the system CT).

(f) The theory should allow for standard interpretations.
An interpretation of the linguistic expressions of the truth theory should be fixed. According to the criterion (a), truth is a predicate which applies to singular terms. The intended interpretation for those syntactic objects is the most natural one: singular terms are intended to refer to sentences and the truth predicate to express the property 'to be true'. In more logical terms, theories of truth not only should have a model (of course they should since they should be consistent), but they should also have a standard model.

(g) The outer logic and the inner logic should coincide.
For a theory of truth T we distinguish *inner (or internal) logic*, the set of sentences that are provably true, i.e. all ϕ such that $\mathsf{T} \vdash T\ulcorner\phi\urcorner$ and *outer (or external) logic*, the set of sentences that are provable,

that is such that $\mathsf{T} \vdash \phi$. In certain systems they coincide, but there are theories in which they do not. For example, the outer logic of KF is classical, while its inner logic is strong Kleene 3-valued logic. This criterion, being an extension of (b) for the whole language, seems very plain as well. Moreover, note that it is entailed by (d).

(h) The outer logic should be classical.
Leitgeb argues that to keep classical logic unchanged and to add a sophisticated axiom system is better then deviating from classical logic and adding a more natural truth system.

How far can we go in forcing those *desiderata*? It is necessary to understand where to place the balance between the chance offered by formal systems of making arbitrary choices and the necessity of maintaining the naturalness of the theory.

I argue that in order to answer this question one has to clarify his approach toward truth theories. Such systems in fact have a sort of intrinsic bivalence as an identifying trait: they arise as objects of a purely philosophical interest but they also turn out to be powerful tools in the field of mathematics and philosophy of mathematics. In both cases 'artificialness' should be viewed as a chance, but (a)–(h), being philosophical requirements, are more or less compelling according to the adopted approach. In both cases, I believe (a)–(c) must be taken for granted.

If one is mainly interesting in features of the base theory and less in properties of the notion of truth, then few other criteria besides (a)–(c), or maybe none of them, are binding.

From a philosophical point of view, the abandonment of the natural language opens new paths: to work in a context where there is nor misunderstanding nor ambiguities, where the language can be microscopically analysed in all its parts is a kind of research which provides useful contributions to the philosophical debate. It is also the most fruitful strategy for testing ideas by bringing them to the extreme possibility and by analysing the consequences. So, the advantage of reasoning in formal terms must be exploited up to the limit: no restrictions should be *a priori* viewed as unacceptable or problematic, in the spirit of the Carnapian saying according to which there are no morals in logic. Nevertheless, one should eventually come back to philosophy again and in order to do so one should try to maintain the intended interpretation: the possibility of going back must be kept open in order to be able to 'read' the results. That is why the point (f) seems indispensable as well and, in my opinion, this might be a not overly restrictive manner to interpret the request of naturalness for truth theories.

Although formalized languages (as PA) and classical logic are far from ordinary natural language and reasoning (there is a significant loss of complexity in moving from one to another!), a basic similarity does remain and it might represent a kind of anchor for the theory; hence to give up (h) is rather problematic. So I tend to localize the arbitrariness in the choice of the basic principles: choosing a list of axioms is already perceived as arbitrary *per se* up to the point that to advocate restrictions in the name of naturalness seems no more justifiable. That is why, I argue that among those criteria (d) and (e) are the less compelling. I believe that it is precisely this combination of similarity and distance from natural language that makes the world of axiomatic theories of truth so attractive: they are both formal tools that allows one to do manipulations, check results, etc. and, at the same time, they are theories in which the philosophical interpretation remains in the background but it is always 'available'.

At any rate, criteria (a)–(h), just as conditions 1.–3., are not independent of each other: more often than not, to drop or modify one of them has repercussions also to the others.

Chapter 2

Comparing axiomatic theories of truth

Truth theories in their classical presentations are well-delimited formal systems, within which many different and interesting studies can be carried out. But another kind of investigation is attractive as well: a metatheoretical and, at the same time, inter-theoretical one. It consists in placing ourselves outside of the theory and looking at it from a distance. The aim is to focus on its placement with respect to other theories (non only theories of truth) in order to pick out bounds, power and other feature of the theory itself. Facing this topic the main problem is to compare truth theories with other theories and this necessarily involves the broad theme of reducibility.

The aim of this chapter is to investigate what happens when truth, axiomatized by certain theories, is subjected to reductions. The starting point will be a reflection on the notion of reduction between theories, in order to isolate aims and tools. Then I shall focus on truth theories and in this respect I shall argue that axiomatic theories of truth reveal a peculiarity with respect to the problem of reducibility. Indeed they lend themselves well to a metatheoretical investigation. Moreover, I believe that this large family of theories might be a interesting ground for the study of the fruitful and philosophically relevant problem of reduction.

A remark of purely terminological character: I use indifferently the words 'theory' and 'formal system'. Clearly there is an abuse here as far as a formal system is a collection of axioms and rules for generating theorems whereas a theory is a set of formulae closed under logical consequence. So, although a theory might be generated by different formal system, in many cases I shall not distinguish between theories and formal systems.

2.1 On reduction

The analysis of the concept of reduction between theories is a broad field, characterized by a jungle of opinions and a fragmented and unsystematic literature. I shall try to untangle this maze by using as guidelines some distinctions that will allow us to orient ourselves. A remark before starting: these distinctions should not be seen as branches stemming from a common root such that better and better define certain classes, it is rather a complex network of similarities and overlapping inclusions.

Reduction in Natural Sciences VS Reduction in Abstract Sciences. [1] The problem of reduction between scientific theories is a widely dealt issue in philosophy of science. It is also very controversial: there is no agreement among philosophers on what is the meaning to be attributed to a reduction in this field, mainly because it involves both problematic concepts as explanation, prediction, approximation and ontological claims. Anyhow, I try to give a very loose (and maybe rough) characterization focusing my attention on purposes, without dealing with the problem of isolating formal conditions and essential features for reduction. Philosophers have differed in what they regard as necessary for something to be a reduction and throughout the twentieth century three models of theoretical reduction have been isolated: the translation model, the derivation model, and the explanation model[2]. Apart from the chosen model, generally, reduction projects in this context have an *explanatory* purpose: when you reduce one theory to another it is required that the descriptions of phenomena and the predictions of the reduced theory are somehow subsumed by those of the latter. With regard to the first model, I consider the positivist idea that through reduction the subjective methods of social sciences should be replaced with objective methods of physical sciences explanatory and not foundational as well: by reducing one theory to another one seeks intersubjectively understandable explanations and predictions, not a sort of justification for the reduced theory. But things still remain complicated since there are different models or frameworks for inter-theoretic explanation[3].

[1] Of course, this requires a distinction between natural sciences and abstract sciences over which I do not linger, let me take the terms in their intuitive meaning: among natural sciences I include biology, physics, chemistry and so on as distinguished from the abstract or theoretical sciences, as mathematics or philosophy.

[2] Historically, the translation model is associated with the early logical positivists Carnap and Neurath, the derivation model with the later logical empiricists Carl Hempel and Ernest Nagel, and the explanation model with John Kemeny and Paul Oppenheim. In the contemporary debate some criticisms and refinements have been made with respect to the original formulations.

[3] For this issue a landmark in literature is the book *The Structure of Science: Problems*

In abstract sciences things change: there are of course problems of explanation, but in this field the purpose of reductions is mainly *foundational*, to the extent that by reducing one theory we try to bring back it to a more justified one or more primitive system. In abstract sciences the philosophical need of justify a theory becomes crucial: the comparison with phenomena, which is usually a discriminating factor, is missing. So, the intent is to isolate a system that is, for whatever reason, more fundamental than another (not needing justification) and the reduction of another theory to that system somehow justifies the former[4]. At any rate, philosophical relevance of reduction is not simply identified with foundational importance: we shall see that for truth theories the aim, although can be to some extent considered philosophical as well, is not foundational.

Global Reduction VS Local Reduction. [5] This distinction is 'transversal' with respect to the previous one: in both natural sciences and mathematics we found local projects of reduction opposing to global programs. In natural science *global reductionism* is the attempt to build the so-called Theory of Everything, an unified system that collects together different phenomena such as waves, elementary particles, multicellular organism up to social groups[6].

A program of this kind is anything but undisputed; indeed it causes skepticism and is subjected to criticisms: antireductionists abandon the idea that various sciences (physics, astronomy, chemistry, biology) can be unified into a single overarching theory and merely pursue *local projects of reduction.*

In mathematics the contrast between global and local reduction has an historical development: in the late nineteenth and early twentieth century, *reductio ad unum* programs arose in foundations of mathematics; then, as it is usually said, Gödel limiting results undermined the overall plant and, accordingly, global reductionism was abandoned in favour of local projects of reduction. Let us further delve the issue without going into the details, trying to focus on reduction. Foundational programs such as logicism, intuitionism and formalism share the belief that mathematics can be brought back to simple contents, (abstract entities which may be captured logically, mental acts or finitistic portion of mathematics). This is a reductionist perspective, and it is

in the Logic of Scientific Explanation by Ernest Nagel (1961). For a survey of the problem of reduction in physics see [2].

[4] See the next distinction for a deeper discussion about this topic.

[5] This distinction has been used by Feferman in order to introduce his proposal, see Feferman [17].

[6] For a historical survey and a discussion of the contemporary key issues of scientific reductionism see [9].

global for the aim is not to reduce one system to another but to reduce the *whole* mathematics to something intuitively justified.

In the early twentieth century these foundational programs ran into several difficulties (foundational crisis of mathematics); for our purposes it is enough to concentrate on results which mark the breakdown of global projects of reduction. The watershed is represented by Gödel's theorems: during the thirties, Gödel proved the incompleteness of formal systems containing a (quite limited) portion of arithmetic. The second incompleteness theorem states the unprovability of the consistency of a (reasonably powerful and recursively axiomatized) mathematical theory within the theory itself[7]. This maybe is the more harmful result for foundational programs *à la* Hilbert. At any rate is not my intention to address the interesting debate about the impact of Gödel's theorems on Hilbert's program, but, for sure, incompleteness theorems determine the need to go outside the boundaries of the formal system itself: the project to reduce all mathematics to a small subpart cannot be achieve. Nevertheless, although Hilbert's solution for foundation of mathematics seemed to be no more defensible in its original formulation, Hilbert's program has not been completely abandoned: it was somehow resumed by Gentzen with the consistency proof of arithmetic[8], Kreisel and his *modified Hilbert's programme*[9] and lastly, by Feferman who promoted a *relativized form of Hilbert's program*[10]. Of course, the monistic view typical of global reductionism has been dropped in favour of an alternative approach with a local character. Local projects of reduction simply consist in reducing somehow a significant formal system to another rather than all formal systems to a single one. In other words, rather than isolating a system (or subsystem) that would serve as 'foundation' for the entire 'mathematical building', one wonders what rests on what in the spirit of the Feferman's works. However, this local perspective does not mark the abandonment of the foundational intent[11]. I refer to the cited sources for further reading on this topic, since my survey is primarily concerned with reduction. Feferman describes the new general pattern of foundation as follows:

> A body of mathematics $\mathfrak{M}$ is represented in a formal system T_1 which is justified by a foundational or conceptual framework $\mathcal{F}_1$. T_1 is reduced proof-theoretically to a system T_2 which is justified by another, more elementary such

[7] For a comprehensive explanation see Boolos and Jeffrey [4], chapters 17 and 18.

[8] Cfr. Gentzen [23].

[9] G. Kreisel, Hilbert's programme, «Dialectica» 12, (1958).

[10] See mainly Feferman [15].

[11] For a survey on different foundational ways see Feferman [?], particularly pp 10–13.

framework $\mathcal{F}_2$[12].

This kind of reduction is, according to Feferman, still foundational but *partial*, for T_1 is just a *part* of what can be justified by $\mathcal{F}_1$. Hence, the role of reductive proof theory becomes predominant.

Theory Reduction VS Ontological Reduction. *Theory reduction* is a inter-theoretic relation that holds between theories when one of them is somehow contained in the other one. There are many nonequivalent definition of "theory S is reducible to theory T". As Niebergall[13] pointed out, they differ from each other to such high degree that only some triviality, like the arity of the relation, will remain. In the next section I am going to talk about the three best known notion of reduction between formal systems: *proof-theoretic reduction*, *relative interpretation*, and *translation*. Whether these notions can be considered reduction in a strict sense is matter of discussion[14].

Whereas theory (or theoretical) reductions are reductions between theories, the notion of *ontological reduction* is used to talk about a relation between phenomena or entities. Ontological reductions are seen as ways to unify and simplify our ontology, against needless multiplication of entities. We can further distinguish between a stronger claim about ontological reduction and a weaker claim. In the first case there is a genuinely ontological attitude towards the existence of the objects involved: objects which are to be reduced are already presupposed as really existent. The question, then, is what does it happen to those entities after an ontological reduction? Are they eliminated or just identified with the reducing ones? Both options seem hard to be maintained especially if one talks about abstract objects such as numbers or sets. On the one hand, there is multiplication of entities as well: even though the objects are reduced, in the sense of identified, however, their existence would be asserted and so the pursued parsimony would be not achieved. On the other hand, how can the effect of a reduction eliminate objects that were previously seen as existent? But there is also a weaker claim according to which what really happens in a reduction between theories is that certain assumptions (axioms or theorems) of the former can be carried out in the language of the latter; so, in a slightly paradoxical way, ontological reduction is a purely syntactical claim.

Homogeneous Reduction VS Inhomogeneous Reduction. This distinction, suggested by Nagel[15], belongs to the field of natural science, but it

[12] Cfr. Feferman [17], p. 73.

[13] Cfr. Niebergall [35], p.27

[14] Hofweber in [30] maintains a polemical position with respect to this claim.

[15] For a taxonomy of inter-theoretical relations see also Sklar [43].

can also be adapted to our purposes. Reductions are distinct depending on whether theories involved share the same conceptual apparatus or not. The first obstacle to clarify is: what do we understand by saying 'concept'? Without committing myself with philosophical claims about the ontological status of concepts, their internal structure or their relationship with language, I just regard as conceptual apparatus the non-logical symbols of the language. Therefore, although I am aware of committing an abuse, with concepts of a certain theory we understand its descriptive apparatus. Moreover, I take descriptive symbols (predicates and individual constants) of the theory language in their formal purely syntactical presentation, without considering the possible interpretations. After this clarification, let us outline the distinction.

Two theories are said homogeneous if they share the same conceptual apparatus, so the *homogeneous reduction* is a reduction in which the concepts of the reduced theory are a subset of those of the reducing theory.

Conversely, when the reduced theory contains some concepts not present in the reducing one we have an *inhomogeneous reduction.* In this case the problem of reduction gets more complicated, for it involves ontological issues.

Before dealing with some open issues concerning reduction, I shall focus on theory reduction and introduce different reductive techniques.

2.2 Notions of reduction between formal systems

I am going to talk about three notions of reduction between formal systems, known in metamathematical literature as *proof-theoretic reduction*, *relative interpretation* and *translation.* A preliminary remark is required in order to avoid ambiguity: whatever relation between theories can be considered as proof-theoretical reduction because it is investigated by using means with a proof-theoretical character. In this sense even relative interpretability is just one special form of proof-theoretical reducibility. I shall use this notion in its strictest sense to be clarified by a definition. This leads to two definitions of reduction which are distinct to such an extent that a system S can be proof-theoretically reducible to T without being relatively interpretable in it, and *vice versa.* In what follows, the theory to be reduced will always be referred to as the source theory (S). The theory to which one is attempting to reduce the source theory will be known as the target theory (T).

Finally, a remark of a purely notational character: considering two theories S and T, if $\mathcal{L}_\mathsf{S}$ and $\mathcal{L}_\mathsf{T}$ are their language respectively, then I write $\mathcal{L}_\mathsf{S} \cap \mathcal{L}_\mathsf{T}$ to express the common part of them and $\mathcal{L}_\mathsf{S} \cup \mathcal{L}_\mathsf{T}$ for their union.

2.2.1 Proof-theoretic reduction

For simplicity, all formal systems considered are assumed to be primitive recursively axiomatized and 'enough powerful', namely they contain a portion of arithmetic (at least the system $\mathrm{I}\Sigma_1$). We deal with primitive recursive representations of their proof predicate and provability predicate. Given S, we write $\mathrm{Proof}_{\mathsf{S}}(y, x)$ to express that y codes a proof in S of the formula coded by x, and $\mathrm{Bew}_{\mathsf{S}}(x)$ for $\exists y(\mathrm{Proof}_{\mathsf{S}}(y, x))$. Furthermore, since proofs can be considered as finite sequences (or trees) of formulae, if p is (the code of) a proof then $\mathrm{End}(p)$ is the end-formula of p. For the sentence $\neg\mathrm{Bew}_{\mathsf{S}}(\ulcorner 0 = 1 \urcorner)$, namely the consistency of S, we write $\mathrm{Cons}_{\mathsf{S}}$.

The idea of proof theoretic reduction of a theory S to another theory T is that we have an effective method, i.e. a primitive recursive function f, for transforming proofs in S of formulae of a set Φ into proofs in T of the same formulae and this is established in a third system W, which has to be included in T.

Definition 2.2.1. Let S, T be theories as stated and p a proof in S. Let, moreover, Φ be any primitive recursive class of formulae contained in $\mathcal{L}_{\mathsf{S}} \cap \mathcal{L}_{\mathsf{T}}$ and defined by the formula $\Phi(x)$.

S is *proof-theoretically reducible to T conservatively for Φ, provably in W*, $\mathsf{S} \leq \mathsf{T}[\Phi]($ in $\mathsf{W})$, if there exists a primitive recursive $f\colon \mathrm{Proof}_{\mathsf{S}} \to \mathrm{Proof}_{\mathsf{T}}$ satisfying:

(i) for each p, ϕ, if $\mathrm{Proof}_{\mathsf{S}}(p, \phi)$ and $\phi \in \Phi$ then $\mathrm{Proof}_{\mathsf{T}}(f(p), \phi)$,

(ii) $W \vdash \forall x, y(\mathrm{Proof}_{\mathsf{S}}(y, x) \wedge \Phi(x) \to \mathrm{Proof}_{\mathsf{T}}(f(y), x))$.

In other words, it is not enough to have an effective map f from proofs in S to proofs in T: it is also essential for the notion of proof-theoretic reduction that this is provable by some restricted means, namely *at least* in the target theory itself or, even better, in a weaker system such as Primitive Recursive Arithmetic PRA, or $\mathrm{I}\Sigma_1$. In general, the cases considered are $\mathsf{W} = \mathrm{I}\Sigma_1$ and $\mathsf{W} = T$ and the matching reducibility relation is said *uniform* or *non uniform*, respectively. When $\mathsf{W} = T$ we say just S is *proof-theoretically reducible* to T for Φ. The more significant are restrictions on the instruments used to carry out a reduction, the more the reduction itself will be meaningful. That is why, in general, the uniform notions are weaker than the uniform ones.

Let us define another inter-theoretic relation which will be very useful for our discussion:

Definition 2.2.2. A theory S is *conservative over T for* Φ if all formulae ϕ of Φ which are provable as theorems in S are also theorems of T.

S is a *conservative extension of T for* Φ if, in addition, S is an extension of T.

The requirement (i) of the definition 2.2.1, once it is assumed as hypotheses, is enough to state the conservativity of S over T for formulae in Φ, i.e.

$$\phi \in \Phi \text{ and } \mathsf{S} \vdash \phi \Rightarrow \mathsf{T} \vdash \phi. \tag{2.1}$$

That is because if S is proof-theoretically reducible to T with respect to a certain class of formulae, then S does not prove anything about those formulae over and above what T proves (and, moreover T knows this). That is, S is conservative over T with respect to a certain class of formulae. The converse would be true if (2.1) is provable in the target theory. Furthermore, if the class Φ contains the equation $0 = 1$ as well, and typically this is the case, then (2.1) yields:

$$\mathsf{T} \text{ consistent } \Rightarrow \mathsf{S} \text{ consistent.}$$

And under the further hypotheses that (ii) holds, we have that in W the relative consistency between S and T is proved. Formally:

Definition 2.2.3. $\mathsf{S} \leq_{RC} \mathsf{T}($ in $\mathsf{W})$, in words: S is *relatively consistent to T, provably in W* if

$$\mathsf{W} \vdash \mathrm{Cons}_{\mathsf{T}} \rightarrow \mathrm{Cons}_{\mathsf{S}}.$$

Again, if $\mathsf{W} = \mathrm{I}\Sigma_1$ this relation is said *uniform*, and *non uniform* if $\mathsf{W} = T$.

Concluding this brief review, it might be interesting to see the model-theoretic counterparts of the notions defined:

Definition 2.2.4. S is *model-theoretically reducible to T conservatively for* Φ, $(\mathsf{S} \trianglelefteq \mathsf{T}[\Phi])$ if there exists a function $f\colon \mathrm{Mod}(\mathsf{T}) \rightarrow \mathrm{Mod}(\mathsf{S})$ such that:

(i) $\forall \mathcal{M}(\mathcal{M} \models \mathsf{T} \Rightarrow f(\mathcal{M}) \models \mathsf{S})$,

(ii) $\forall \mathcal{M} \forall \phi \in \Phi(f(\mathcal{M}) \models \phi \Rightarrow \mathcal{M} \models \phi)$.

Definition 2.2.5. A theory S is *semantically conservative over T* if every model of T can be expanded to a model of S.

Model-theoretic reduction implies conservativeness as well, but it is important to point out that the syntactical conservativity and the semantical one do not always agree.

2.2.2 Relative interpretation

For the first time the precise notion of *relative interpretability* has been defined by Tarski. In [48], he employed this kind of notion as an indirect method to establish whether a formalized theory is decidable or not, by reducing it to another theory for which the decision problem has already been solved.

Intuitively, S is relatively interpretable in T if for each relation, function and constant symbol of the language $\mathcal{L}_{\mathsf{S}}$ of S there is a *possible definition* of it in $\mathcal{L}_{\mathsf{T}}$. Then, with each formula of $\mathcal{L}_{\mathsf{S}}$ it is associated as its interpretation in $\mathcal{L}_{\mathsf{T}}$ a formula obtained by substituting the respective definitions for non-logical symbols and by relativizing all quantifiers. Before giving a more technical definition, we have to answer some questions like: what is a possible definition of a given symbol in a theory to which it does not belong? What does the relativization of quantifiers mean and why is it important?

A possible definition for a predicate P with arity n of $\mathcal{L}_{\mathsf{S}}$ in the language of T is a formula of the following form:

$$\forall \overrightarrow{x}(P(\overrightarrow{x}) \leftrightarrow \phi(\overrightarrow{x})), \tag{2.2}$$

where $\phi(x_1, \ldots, x_n)$ is a formula in the language $\mathcal{L}_{\mathsf{T}}$, containing no more than n free variables. The sentence (2.2) is not a sentence of $\mathcal{L}_{\mathsf{T}}$, as the symbol P^n does not belong to $\mathcal{L}_{\mathsf{T}}$; but it is a sentence in every extension of T containing the predicate symbol P^n. The same happens for all predicative constant in $\mathcal{L}_{\mathsf{S}}$. We consider the identity predicate, $=$, as a logical symbol, so it is always translated by itself.

A possible definition of an n-ary function symbol in T is:

$$\forall \overrightarrow{x} \forall y(f(\overrightarrow{x}) = y \leftrightarrow \phi(\overrightarrow{x}, y)). \tag{2.3}$$

where $\phi(\overrightarrow{x}, y)$ is a formula of T with exactly the variables $x_1, \ldots, x_n, y$ free. Moreover, this formula cannot be arbitrary: it must be functional and, so, T must prove:

$$\forall \overrightarrow{x} \forall v \forall u(\phi(\overrightarrow{x}, u) \wedge \phi(\overrightarrow{x}, v) \rightarrow u = v).$$

However, we can assume for simplification that our theories are formulated in a purely relational language, i.e. a language which contains only relation signs.

Individual constants are treated as 0-place function symbols. Thus, if c is a constant of S, its definition in $\mathcal{L}_{\mathsf{T}}$ is:

$$\forall y(c = y \leftrightarrow \phi(y))$$

with $\phi(x)$ formula of $\mathcal{L}_{\mathsf{T}}$ and such as in T is provable the following sentence:

$$\forall u \forall v(\phi(u) \wedge \phi(v) \leftrightarrow u = v).$$

In order to relativize the quantifier we take a symbol δ not in the language of the source theory S and then we substitute all subformulae $\forall x \psi(x)$ and $\exists x \psi(x)$ in formulae of S with $\forall x(\delta(x) \rightarrow \psi(x))$ and $\exists x(\delta(x) \wedge \psi(x))$, respectively. Generally $\delta(x)$ is a formula of the target theory T, which provides a defined range of variation for the variables of $\mathcal{L}_{\mathsf{S}}$.

Let us see with an example taken from Niebergall [35], why this operation is fundamental. Consider the reduction of the Peano Arithmetic (PA) to

Zermelo-Fraenkel set theory (ZF). We have possible definitions, like $0 := \emptyset$ and $s(x) := x \cup \{x\}$, for the constants and functions of PA. But if we translate a sentence of $\mathcal{L}_{\mathsf{PA}}$ which is provable in PA into a sentence of $\mathcal{L}_{\mathsf{ZF}}$, simply interchanging *definiens* and *definiendum*, we obtain a sentence which is not provable in ZF:

$$\mathsf{PA} \vdash \forall x(x \neq 0 \rightarrow \exists y(x = s(y))), \tag{2.4a}$$

$$\mathsf{ZF} \nvdash \forall x(x \neq \emptyset \rightarrow \exists y(x = y \cup \{y\})). \tag{2.4b}$$

Reducing PA to ZF, we want to translate whatever is PA-provable about PA-objects into something ZF-provable about the matching ZF-objects. In (2.4a) variables range over the natural numbers, while in (2.4b) over arbitrary sets, including the first infinite ordinal ω which is different from the empty set but not a successor ordinal. So, a real correspondence between PA- and ZF-objects must associate all natural numbers not with all sets, but only with all finite ordinals. Therefore, an adequate set-theoretical version of (2.4a) is provided by restricting in (2.4b) the quantifiers to ω:

$$ZF \vdash \forall x \in \omega(x \neq \emptyset \rightarrow \exists y(x = y \cup \{y\})),$$

where $x \in \omega$ stands for the formula in $\mathcal{L}_{\mathsf{ZF}}$ expressing that x is a finite ordinal.

After these explanations, we can give a formal definition of the notion of relative interpretability, following Feferman[16]:

Definition 2.2.6. Let S, T be theories in finite relational languages $\mathcal{L}_{\mathsf{S}}$ and $\mathcal{L}_{\mathsf{T}}$. Assume that for each k-place relation sign R in $\mathcal{L}_{\mathsf{S}}$ there is a possible definition φ_R in $\mathcal{L}_{\mathsf{T}}$. And let δ be a fixed 1-place formula in $\mathcal{L}_{\mathsf{T}}$ different from all the φ_R.

ι is a *relative interpretation* of S in T with respect to δ if and only if:

(i) $\iota\colon \mathcal{L}_{\mathsf{S}} \rightarrow \mathcal{L}_{\mathsf{T}}$ is primitive recursive,

(ii) for all n, m; $\iota(x_n = x_m) = (x_n = x_m)$,

(iii) for each k-place relation sign R in $\mathcal{L}_{\mathsf{S}}$; $\iota(R(x_1, \dots, x_k)) = \varphi_R(x_1, \dots, x_k)$,

(iv) for all formulae χ, ψ in $\mathcal{L}_{\mathsf{S}}$;

$$\begin{gathered}\iota(\neg\chi) = \neg\iota(\chi),\\ \iota(\chi \vee \psi) = \iota(\chi) \vee \iota(\psi),\\ \iota(\chi \wedge \psi) = \iota(\chi) \wedge \iota(\psi),\\ \iota(\chi \rightarrow \psi) = \iota(\chi) \rightarrow \iota(\psi),\end{gathered}$$

[16] Cfr. Feferman [11], p. 49. This definition describes the same relation introduced by Tarski in [48].

(v) for all formulae ψ in $\mathcal{L}_{\mathsf{S}}$ and all variables x:

$$\iota(\forall x \psi(x)) = \forall x(\delta(x) \to \psi(x)),$$

(vi) $\mathsf{T} \vdash \exists x \delta(x)$,

(vii) for all formulae (sentences) ψ in $\mathcal{L}_{\mathsf{S}}$, if $\mathsf{S} \vdash \psi$, then $\mathsf{T} \vdash \iota(\psi)$.

S is *relatively interpretable* in T ($\mathsf{S} \preceq \mathsf{T}$) if and only if there are a function ι and a formula δ such that ι is a relative interpretation of S in T with respect to δ.

There is, as important variant of this definition, the interpretability *simpliciter* of S in T, i.e. interpretability where the relativizing formula δ is universally valid in T. An interpretability of this kind is equivalent to S being a subtheory of a definitional extension of T, i.e. a theory obtained by the addition of explicit definitions of the non-logical symbols of $\mathcal{L}_{\mathsf{S}}$. As other variant we can imagine an interpretation for which condition (vii) is valid also in the reverse direction. Specifying these intuitive notions of restricted forms of interpretation, we obtain the following:

Definition 2.2.7. Let S, T be theories in $\mathcal{L}_{\mathsf{S}}$ and $\mathcal{L}_{\mathsf{T}}$ and ι a relative interpretation of S in T w.r.t. δ:
ι is *unrestricted* iff $\delta(x) = x$.
ι is *faithful* iff $\mathsf{T} \vdash \iota(\psi) \Rightarrow \mathsf{S} \vdash \psi$.

As well as for proof-theoretic reducibility, even for relative interpretability there is a semantical counterpart. Actually, there are many different model-theoretic definitions of reducibility which are equivalent to relative interpretability, for example the following:

Definition 2.2.8. S is (semantically) interpretable in T if there is a function ι from $\mathcal{L}_{\mathsf{S}}$ to $\mathcal{L}_{\mathsf{T}}$ commuting with connectives and quantifier (which may have to be relativized) such that

$$\forall \mathcal{A} \models \mathsf{T}\ \exists \mathcal{B} \models \mathsf{S}\ \forall \psi \in \mathcal{L}_{\mathsf{T}} (\mathcal{B} \models \psi \Leftrightarrow \mathcal{A} \models \iota(\psi)).$$

Relative interpretation satisfies certain intuitions which can be used, as Niebergall suggested, as *adequacy conditions.* For example, an adequate notion of reducibility should subsume the subtheory relation and, at the same time, should be wider than it. Another property which might be desirable to hold is transitivity. For both the requests, this is the case:

Remark 2.2.1. If R, S and T are axiomatic systems:

(i) $\mathsf{S} \subseteq \mathsf{T}$ implies $\mathsf{S} \preceq \mathsf{T}$ but in general the opposite direction does not hold,

(ii) if $\mathcal{R} \preceq \mathsf{S}$ and $\mathsf{S} \preceq \mathsf{T}$, then $\mathsf{R} \preceq \mathsf{T}$.

Interpretability is also an important tool in proofs of (relative) consistency and decidability. As Visser[17] remarked, interpretations have some very useful preservation properties.

Consistency. Interpretability preserves consistency from target theory to interpreted theory and inconsistency in the reverse direction.

Reflexivity. Mutual interpretability preserves reflexivity.

Decidability. Interpretability preserves essential undecidability from interpreted theory to interpreting theory. Faithful interpretability preserves decidability in the same direction.

Definition 2.2.9. A consistent theory is said to be *essentially undecidable* if it has the property that every consistent extension is undecidable.

2.2.3 Translation

There are different definitions of what constitutes a translation; but the minimal assumptions that a function f has to met in order to be a translation are, as relative interpretation:

(i) $f\colon \mathcal{L}_{\mathsf{S}} \to \mathcal{L}_{\mathsf{T}}$ is primitive recursive,

(ii) if $\mathsf{S} \vdash \varphi$, then $\mathsf{T} \vdash f(\varphi)$,

(iii) f preserves propositional operations[18].

Therefore we define:

Definition 2.2.10. $\mathsf{S} \leq \mathsf{T}$ holds in the sense of translation, if there is a function f satisfying (i), (ii), (iii).

Just as for relative interpretation, we have:

$$\mathsf{S} \leq \mathsf{T} \wedge \mathsf{T} \text{ consistent } \Rightarrow \mathsf{S} \text{ consistent.}$$

However, we are going to compare theories which share their base theory and differ only in the truth axioms. For this reason, in order to define translation functions between two truth theories S and T it will be enough

[17] Cfr. Visser [49], p. 5.

[18] Note that this assumption can be further restricted:

(iii') $f(\neg\varphi) = \neg f(\varphi)$.

In this way, the definition of translation is more general since it can be applied to cases in which the two theories are formulated in different logical systems (for example if the logic of S is classical and that of T is intuitionistic). For our purposes, this generalization is not required because we shall deal just with systems formulated in classical logic.

to show whether the former can translate the truth predicate of the latter. As we shall see, in doing this we shall use the recursion theorem. This theorem guarantees the existence of functions built by recursion. Why do we need a recursive definition of the translation? Suppose that a translation substitutes a truth predicate T with another formula θ; this function should substitute all occurrences of T, even the ones in which T is just mentioned for example in the formula $T(\underset{\cdot}{T}t)$ both of the occurrences of T should be substituted. To this aim, in general, if f is our translation function substituting T with θ, then for any term s the translation $f(T(s))$ should be $\theta(\underset{\cdot}{f}(s))$ and not merely $\theta(s)$, where $\underset{\cdot}{f}$ is the formula representing f in the language of arithmetic. Therefore the translation function f is recursively defined in terms of f itself and $\underset{\cdot}{f}$. The existence of such a function is guaranteed by the recursion theorem.

2.3 Open issues about reduction

Investigating formal systems we are concerned in finding out their proof-theoretical strength, expressive power and ontological commitments. Moreover, we would like to prove other properties of systems such as decidability, consistency and so on. Reductions and, in general, intertheoretic relations are an essential tool for these purposes.

In the previous section we have seen different candidates for a general explanation of "theory S is reducible to theory T". The further step is to outline adequacy criteria for suggested explications of 'reducibility'. Niebergall in [35] tried to trace some conditions. First of all a non trivialization of the relation is required, that is to say not every theory should be reducible to every theory[19]. Moreover, reducibility should be a proper weakening of the subtheory relation: if S is subtheory of T, then S should be reducible to T, but the reverse, in general, should not hold. Having said this, a reducibility relation ρ must meet the following:

1. $\mathsf{S} \subseteq \mathsf{T} \Rightarrow \mathsf{S}\rho\mathsf{T}$.

2. $\mathsf{S}\rho\mathsf{T}$ and $\mathsf{T}\rho\mathsf{R} \Rightarrow \mathsf{S}\rho\mathsf{R}$.

3. If $\mathsf{S}\rho\mathsf{T}$ then $\mathrm{Cons}_{\mathsf{T}} \Rightarrow \mathrm{Cons}_{\mathsf{S}}$.

4. Let $\bar{A}$ be the deductive closure of A in classical first order logic: if $\mathsf{S}\rho\mathsf{T}$ then $\forall\phi \in \mathsf{S}\ \exists\psi \in \mathsf{T}(\{\bar{\phi}\}\rho\{\bar{\psi}\})$.

5. For S and T finitely axiomatizable, $\mathsf{S}\rho\mathsf{T}$ implies $\mathrm{I}\Sigma_1 \vdash \mathrm{Cons}_{\mathsf{T}} \rightarrow \mathrm{Cons}_{\mathsf{S}}$.

[19] In a mathematical context this requirement is formulated as follows: no every recursively enumerable theory is reducible to Q or $\mathrm{I}\Sigma_1$ (weak subtheories of PA); and $Th(\mathbb{N})$ is not reducible to Q or $\mathrm{I}\Sigma_1$, where $Th(\mathbb{N})$ is the set of sentences of $\mathcal{L}_{\mathsf{PA}}$ holding in the standard model of natural numbers.

Niebergall then turned these clauses into axioms in order to provide an axiom system for reducibility[20]; nevertheless, apart from this attempt they could equally well be considered as loose guidelines for an inquiry about reduction.

Theory reductions, especially between mathematical theories, are intensively studied among logicians and philosophers. In what follows I summarize some open issues concerning them:

1. Although Niebergall's criteria are strong enough to reject some alternative definitions of reducibility, they are non conclusive with respect to the dispute about the individuation of the prime reducibility concept between relative interpretability and proof-theoretic reducibility. Since reasons for both the stances are plainly formulated by their supporters, I refer to their papers: Feferman [17], Appendix section A.6, and Niebergall [35], sections 2 and 4.

2. Another issue concerning reductions, or better technical results in reductive proof theory, is their pretended philosophical relevance, raised by Hofweber [30]. In the domain of mathematics, he distinguished between foundational importance and large-scale philosophical importance of reduction. The former concerns questions about axioms, strength and so on, while the latter questions about mathematical objects, knowledge and such. He argues that neither relative interpretation nor proof-theoretic reduction are sufficient for a philosophically relevant reduction, for they both are relations between formal systems and have no impact on underlying forms of reasoning, i.e. the interpreted or the informal theories that are modeled by those systems. We shall see that talking about truth theories this gap is weakened and somehow disappear.

3. Another controversial issue is the relationship between theoretical and ontological reduction: from the fact that a theory is reducible to another — in one of the senses previously outlined — does it follow that the objects which one theory talks about are the same as the objects of the other theory? For example can the relative interpretability of PA in ZF be seen as an argument in support of the claim that natural numbers are sets? And moreover, which notion of theory reduction, if there is any, involves an ontological reduction between objects of those theories? There are different views. According to Niebergall, although the notion of theoretical reducibility between S and T is a necessary condition in order to have an ontological reduction — or at least it should be entailed from an ontological reduction —, it does not seem enough: there must be a further unspecified condition between theories. He warns against the risk of connecting purely syntactical

[20] See Niebergall [35], especially the third section.

relations with statements about ontological assumptions. Hofweber shares the same view: from a proof-theoretic reduction does not follow that involved theories talk about the same things. A position diverging from these is due to Halbach. He sees ontological reduction in the weaker sense, that is as a mapping between assumptions of different nature (arithmetical, set-theoretical, truth-theoretical etc.), more than a mapping between different kind of objects (set, number etc.). Then, relative interpretations are examples of ontological reduction, once it is understood in this way.

Since discussing these questions without specifying a domain is not advisable, in the next section I shall attempt to study reduction having as area of application axiomatic theories of truth. My purpose is both to understand the relevance of a metatheoretical analysis for truth theories and to shed light on problems relating to reduction in general. In particular, about the aforementioned issues we shall see to what extent for reduction between theories of truth we are led to accept a methodological pluralism and philosophical significance of results. Finally, I shall discuss the ontological problem.

2.4 Reduction and truth

What I plan to do now is to put together previous considerations, by approaching a general issue:

What does it happen if we apply the notions of reduction to truth theories?

The answer must be articulated from different perspectives:

- 'justificational' perspective: why comparing theories of truth? To what extent the act to compare truth theories can tell us something about the notion of truth?

- methodological perspective: how to compare truth theories? How to calibrate our instruments to ensure they fit for our purposes?

- technical perspective: I argue that other attitudes must be set aside for a while in order to give way to logic.

- philosophical perspective: what is the *meaning* of the notions of reduction when applied to this context? Namely, how they should be interpreted? What do they entail? Which are their side effects on conceptual aspects underlying truth theories?

2.4.1 Why reduction in axiomatic theories of truth

Arguably, a justification in this frame has to be done *a posteriori*, but of course we can ask what we expect from an investigation of this kind and which are our aims. Literature abounds in different views about the purposes, but they can be collected into two different approaches to theories of truth (which are clearly not mutually exclusive): an instrumentalist approach and a philosophical or conceptual one. I argue that whatever is the attitude, a metatheoretical investigation allows us to achieve some important results.

On the one hand, truth theories can be studied with philosophical interest and, since every theory embodies a different view, the act to compare them is a way to learn something more about the underlying stance on truth. On the other hand, the aim might be to use them as a tool or a device to increase the power of other theories and so what does really matter is their proof-theoretical power. And a practicable strategy is to establish the proof-theoretical strength of a theory by reducing it to another one whose power is well known. Therefore, a metatheoretical inquiry seems to be an useful tool in dealing with truth, but let us tackle this problem in details.

To what extent comparing truth theory is philosophically relevant? Of course the answer requires the abandonment of a foundational perspective, in the sense stated before: the aim is no more to lead back a theory to another theory which is somehow more fundamental in order to justify it, indeed, in some ways, we witness its reversal. There are at least two problems with respect to a foundational claim. First, it seems hard to find 'criteria of justification', that is to say something that makes a theory more justified than the others. Such a criterion might be the correspondence with an immediate insight or essential features about truth. But axiomatic theories of truth can be seen as 'thought experiments' and each way must be considered equally interesting and desirable in principle. Otherwise the peculiar gain of axiomatizations is lost, namely the fact that it allows one to employ any kind of concepts, even the less intuitive, and to explore (formally) consequences of their employment. Secondly, the reducing theory, rather the reduced one, gets more benefits. As a very simple example consider the fact that in $\mathsf{CT}{\upharpoonright}$ typed Tarski-biconditionals can be derived, that is the theories $\mathsf{TB}{\upharpoonright}$ and $\mathsf{UTB}{\upharpoonright}$ are subtheories of $\mathsf{CT}{\upharpoonright}$[21]. This reduction makes us aware of the expressive power of $\mathsf{CT}{\upharpoonright}$, to some extent is a mark in its favour and in no way is a justification for the reduced theory. So, the more notions of truth are 'simulated' in a theory, the more it becomes attractive.

This brings us to a crucial point: the abandonment of foundational perspective does not entail a loss of philosophical relevance for the problem of reduction and truth. Even better, it is exactly with respect of this issue that theories of truth reveal their own peculiarity. We are dealing with syntacti-

[21] For a proof see Halbach [28], p. 66.

cal theories, which are to some extent *in vitro* settings suitable to study the behaviour of the truth predicate. The loaded question about (the essence of) truth is defused by turning it in a harmless and, maybe, more productive one: how does the predicate 'to be true' work? In other words, the issue is operational, with Wittgensteinian echoes: the meaning of T is fixed by providing axioms and rules. Clearly, each axiomatization is built in order to reflect a particular conception of truth. It is interesting, from a philosophical point of view, to investigate the links between different claims about truth and this can be done by comparing theories that somehow 'determine' them. In other words, I argue that reductions between axiomatic theories of truth have a philosophical relevance because of a peculiar feature: a deep connection between axiomatic framework and what the theories are about. The formal apparatus in a theory of truth is something that cannot be isolated at all. So, reduction between truth theories is neither just matter of words nor just "an association of the relevant formal languages in the right way"[22], rather it should be considered as an interesting and philosophically relevant reduction.

Another reason to pursue reduction projects is the following: reductions provide a relative consistency proof. As seen before, whatever notion of reduction is adopted, if a theory S is reducible to another theory T which is known to be consistent, then S is consistent. The claim that consistency proofs represent a philosophically relevant reason to compare truth theories might be further questioned from two different points of view:

- why theories of truth should be consistent?
- why should we, as philosophers, care about having a consistency proof?

Consistency for an *axiomatic* theory of truth is everything but a side issue, it is not just a nice or auspicabile feature, but something more: in my opinion, it should be considered as a necessary requirement. When paradoxes like the Liar can be carried out in a theory of truth, they threaten the whole system and the motivations that underlie it. For this reason all efforts are directed to avoid a contradiction, and in particular to ensure that the Liar is not derivable.

Inconsistency for a formal system S can be understood in two ways:

[Ctr1] Derivability of a contradiction, that is $\mathsf{S} \vdash \bot$.

[Ctr2] Derivability of each sentence, that is for all ϕ $\mathsf{S} \vdash \phi$.

It is well known that [Ctr1] and [Ctr2], in an intuitionistic framework (and *a fortiori* in a classical one), entail one another. Of course inconsistency of the second kind is something we would like to avoid, as otherwise we would

[22] Cfr. Hofweber [30], p. 132.

have a trivialization of the theory. By choosing a suitable logic[23] one can *defuse* the implication [Ctr1] → [Ctr2], suggestively called *explosion.* Then, an escape route for [Ctr1] is simply to accept that contradictions are provable (by denying *Law of Non-Contradiction*): some sentences — called *dialetheias* — fall into *truth-value gluts*, that is, are both true and false[24]. Since our theories extend PA, which is known to be consistent, those sentences should belong to $\mathcal{L}_T$. Typically, a *dialetheia* is the Liar sentence, λ. From a semantical point of view, λ is allowed to be both true and false, that is it belongs both to the extension and the antiextension of the truth predicate. At any rate, in these cases inconsistency is avoided by deviating from classical logic. Regardless of justifiability of this kind of approach[25], resultant theories of truth are still consistent. But can the claim of consistency be completely dropped in an axiomatic approach to truth? It seems that the expediency of an inconsistent axiomatic theory of truth is really hard to maintain.

For what concerns the second issue, Hofweber wonders whether we, as philosophers, should care about having a consistency proof. Note that he is talking about mathematical theories:

> It might seem that the answer why we would want to have a consistency proof is obvious: we want to be sure that a certain theory is consistent, so to be sure that no paradoxes or contradictions can be derived from certain axioms. This, of course, would be nice to know. But it is at first not so clear why knowing this, or being able to prove this, should be taken to be of much philosophical significance. After all, a consistency proof shows that the axioms of a certain axiomatization of a branch of mathematics are consistent. It does not thereby necessarily show anything about the branch of mathematics, only about the axiomatization of the reasoning within that branch of mathematics[26].

Now we ask: what about truth theories? Even supposing that for mathematic this is so — and this claim can be further questioned — does the same irrelevance hold for truth theories? Or, rather, does the consistency show something about truth? I implicitly accept, and it would be difficult to argue otherwise, that axiomatic theories of truth *per se* have a philosophical relevance. Since I said we seek a consistent axiomatization, consistency proof automatically gets the same relevance. Moreover, a peculiarity of theories of truth is a close connection between axioms and what theories are about. This makes consistency proofs very important from a philosophical point of view. However, reductions are not the only tool to this aim: consistency

[23] A paraconsistent logic, for a philosophical introduction see Priest [38], for a more technical one see Priest and Tanaka [39].

[24] For a defense of this approach see Field [19].

[25] For a negative opinion see Leitgeb [33], criterion (h).

[26] Cfr. Hofweber [30], p. 138.

proofs can as well be provided by using model-theoretic means, nevertheless as Halbach[27] notes, proof-theoretic reductions are preferable as they are more parsimonious in terms of resources. So, having a syntactical consistency proof is always desirable. Concluding this remark, I hope I have provided a further answer to the opening question: why reduction in axiomatic theories of truth?

Furthermore, Halbach points out that reductions can be seen as *completeness proofs* as well, where completeness is taken in a weak sense that I shall explain. Given a semantical construction (e.g. Kripke's fixed point theory), different axiomatizations can be setting up to capture it. We would like to verify whether a set of axioms is an axiomatization of a particular semantical theory as complete as possible.

> As pointed out, 'completeness' cannot mean that all sentences valid in a semantic construction are provable. But we may show that the system is complete by showing that the tools employed in the semantical construction are available in the system[28].

A semantical construction is carried out in informal mathematics, but in most cases certain second-order systems are sufficient. Reduction of those second-order principle to truth-theoretic principles might be seen as an adequacy condition for the axiomatization.

These reasons justify a philosophical interest towards comparative study of different theories of truth and show how, even in issues of a conceptual nature, reductions can be good tools.

Additionally, theories of truth can be treated instrumentally. Truth predicate can be added to mathematical theories in order to increase their expressive (and proof-theoretical) power[29] and also, we shall see, in order to pursue an ontological parsimony. Following this approach, it is even more immediate to see the utility of reduction. First, it can be shown the relative interpretability of a theory of truth in an arithmetical theory, more exactly some subsystems of second order arithmetic define the truth predicate of certain theories. Conversely, wide parts of mathematics can be developed in truth systems. Previous results taken together give us respectively the upper and the lower bound of proof-theoretical strength of truth theories. In other words, by proving an equivalence between subsystem of second order arithmetic and theories of truth we get a measure of the power of the theory itself. To be aware of this strength is essential when the perspective is to use the truth predicate as a tool to increase the expressive power of mathematical theories. Therefore, even (and maybe mainly) from an instrumentalistic

[27] Cfr. Halbach [26], p. 98

[28] Ivi, p. 99.

[29] Since the increase might be substantial, reductions of subsystems of second order arithmetic to theories of truth contribute to feed the philosophical debate about deflationism.

point of view metatheoretical investigations turn out to be indispensable.

2.4.2 Methodological perspective

Supposing we have provided enough reasons to carry out a metatheoretical investigation, we have now to face the issue from a methodological point of view. It has been already said that the strategy is comparing axiomatic theories of truth, but comparing with what? By means of what? Depending on the purposes of the comparison theories can be compared with:

1. **The base theory.** As said before, we take into account only truth theories that share their base theory. So, throughout this discussion the base theory will be PA. Comparing a truth theory extending PA with PA itself involves two different notions: *conservativity over* PA and *interpretability in* PA[30]. This kind of comparison comes from the need to understand to what extent the truth predicate (i.e. the set of truth axioms) changes the base theory. The former concerns strength and investigates how much the proof-theoretical power of PA increases when the truth predicate is added. Of course, in order to do this one can restrict to consider the arithmetical content, that is the set of theorems which do not contain truth predicate. A survey of this kind leads to not at all trivial results and, moreover, is rich in philosophical implications. For what concerns the latter, we do not care how much stronger the base theory gets, but we wonder whether the theory of truth deviates so much from the base theory not to be more interpretable in it. So, we look at the truth-theoretical content.

2. **Second-order systems.** Theories of truth has been compared with subsystem of second-order arithmetic[31] and many results have been obtained in terms of equivalence[32].

   Second-order arithmetic is formulated in a two-sorted language with one sort of variables $x, y, z, \ldots$ ranging over natural numbers and the other sort $X, Y, Z, \ldots$ ranging over sets of natural numbers, that is subset of $\omega = \{0, 1, 2 \ldots\}$. The language $\mathcal{L}_2$ of second-order arithmetic contains the symbols of PA, and in addition has a binary relation symbol $\in$ for elementhood. Other clauses have to be added to the inductive definition of formulae in the obvious way: there are new atoms of the form $t \in X$ and formulae are closed under second-order quantifiers. As

[30] For a characterization of these notion see the previous section.

[31] Sometimes second-order arithmetic is called ‘analysis’ because it is possible to formalize the real numbers in it, for real numbers can be represented as sets of natural numbers and second order arithmetic allows quantification over such sets. For an overview see Simpson [42], especially Chapter 1.

[32] For a list of result in such field see Halbach [26].

axioms it contains the axioms of PA, induction axiom:

$$\forall X(0 \in X \land \forall x(x \in X \to s(x) \in X) \to \forall x(x \in X)),$$

and comprehension schema:

$$\exists X \forall u(u \in X \leftrightarrow \phi(u)),$$

where $\phi(u)$ is any formula of $\mathcal{L}_2$ in which X does not occur freely. Subsystems are formulated in the same language and have as axioms a proper set of the theorems of second-order arithmetic. Means adopted in comparing such systems with theories of truth are the methods of reductive proof theory like ordinal analysis and relative interpretations.

3. **Other theories of truth.** One can compare truth theories to each other by different means according to the purposes. Fujimoto ([22]) further refined the notion of relative interpretability and proposed a stricter notion of interpretations, *relative truth definability* that leaves the arithmetical vocabulary unchanged and does not relativize the quantifiers of the source theory. That is, just the truth-theoretical content falls under its scope. This feature makes relative truth definability an excellent candidate for the comparison between theories of truth.

   Definition 2.4.1. Let Q and S be theories of truth over languages $\mathcal{L}_{\mathsf{Q}}$ and $\mathcal{L}_{\mathsf{S}}$ respectively, and let $\mathcal{L}_{\mathsf{Q}}$ be $\mathcal{L}_{PA} \cup \{T_i\}_{i \in I}$ where I is a certain index set. The base theory is PA, with the language $\mathcal{L}$. Given a formula θ_i of $\mathcal{L}_{\mathsf{S}}$ for each $i \in I$, a translation $\mathcal{T}_{\vec{\theta}}$ from $\mathcal{L}_{\mathsf{Q}}$ to $\mathcal{L}_{\mathsf{S}}$ is defined as follows:

$$\mathcal{T}_{\vec{\theta}}(\phi) := \begin{cases} \phi, & \text{if } \phi \in \mathrm{AtFml}_{\mathcal{L}}; \\ \theta_i(x), & \text{if } \phi = T_i(x); \\ \neg\mathcal{T}_{\vec{\theta}}(\psi), & \text{if } \phi = \neg\psi; \\ \mathcal{T}_{\vec{\theta}}(\psi_0) \lor \mathcal{T}_{\vec{\theta}}(\psi_1), & \text{if } \phi = \psi_0 \lor \psi_1; \\ \forall x \mathcal{T}_{\vec{\theta}}(\psi), & \text{if } \phi = \forall x\psi; \end{cases}$$

   We say Q is (*relatively*) *truth definable* in S when there exists formulae $\theta_i(x)$ of $\mathcal{L}_{\mathsf{S}}$ for each $i \in I$ such that

$$\mathsf{Q} \vdash \phi \Rightarrow \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi) \quad \text{for all } \phi \in \mathcal{L}_{\mathsf{S}}.$$

   In other words, S defines the truth predicate of Q if and only if there is a monadic formula of the language of the former such that the result of uniforming substituting it for the truth predicate T in a theorem of Q is a theorem of S. Somehow, in S there is a formula that deputizes the truth predicate of Q. This definition is general enough to encompass

> truth theories with more than one truth predicate, such as RT and it can be further generalized in order to embrace cases in which theories do not share their base theory. Additionally, the relation of truth definability meets Niebergall's conditions[33].

Thus, truth theories can be compared from various points of view and means adopted must be the most suitable for the intents. I argue that comparisons must be performed in the pursuit of methodological liberality: one cannot rule anything out *a priori* or underestimate the contribution of any strategy. In the domain of theories of truth emerges the absence of a privileged instrument. This is due to their twofold nature: their are both mathematically useful instrument and philosophically relevant theories. Moreover, the suppleness of the notions of reduction already guarantees a sort of *methodological pluralism*, that is to say that philosophical claims can be investigating by using proof-theoretic means or, *vice versa*, by using purely conceptual comparison tools one can achive strictly proof-theoretical results. The proof that we shall see in the next chapter is an example of this idea.

2.4.3 Logical results and philosophical perspective

A further problem is how to assess the contribution of logic in this field. The claim that the issue should benefit from formal results is questionable as well. There is no agreement about what a result in philosophy is or should be and, furthermore, about the evaluation (positive, negative, neutral?) of the role of logical results (e.g. a meta-theorem). Roughly speaking, given a philosophical debate a technical result might have a negative role if it helps to discredit a philosophical stance (a very classical example: Gödel's theorems with respect to Hilbert's original program.); otherwise, if it contributes to defend certain claim it would have a positive role. I advocate the importance of a dialogue between philosophical claims and technical results in the spirit of mathematical philosophy. This is worth generally speaking, but since we have already chosen an axiomatic approach to truth it becomes essential: being able to benefit from formal results can be considered the main advantage of axiomatic approach towards truth. Hence, I argue that philosophical issues should benefit from formal results, but on the other hand these results have to be found in a field as possible free from philosophical 'contaminations' and restrictions. As said before, once we chose to deal with logical objects (such as formal systems, translations etc.) all strategies, even the more counterintuitive ones, must be pursued. Philosophical ties must be recovered only subsequently in the interpretation of results.

Clearly, the hardest thing is to understand to what degree a result (once identified as negative or positive) is conclusive or decisive with respect to the problem itself. My stance with respect to philosophical interpretation

[33] See section 2.3 on page 41.

of results is the following: it must be *a posteriori* analysis and, above all, the survey must be conducted case by case. This is particularly true for the interpretation of reduction results. It should be made retrospectively: what matters is not just the result but the way taken to reach it, so that even negative results (non-conservativity or non-interpretability ones) can be used as opportunities to learn something more about theories.

Notions of reduction *per se* do not have a clear or univocal content; philosophical implications, if there are any, must be discussed with a look at single cases, for their relevance and meaning are often controversial. As an instance, consider the problem of how to interpret, philosophically speaking, the conservativity of a truth theory over its base theory. This is a crucial point in the modern debate about truth for its close link with deflationist accounts of truth and the interpretation of results is anything but obvious.

2.4.4 Ontological reduction

After outlining the problem of applying reduction to the field of truth theories (with an inquiry about motivations, methods and possible outcomes), we focus on the specificity of axiomatic theories of truth with respect to ontological reduction. I argue the peculiarity of truth theories is a sort of 'adherence' between the theory and its object.

What are the object of a truth theory? Without going deeper into the issue about their nature, in the first chapter I have assumed that the truth bearers are sentences or, more cautiously, something with the same structure of sentences, that is with atomic expressions and logical operations for forming complex expressions. Clearly, sentences are always sentences of some language, therefore in the proposed theories the objects are sentences of the language of the base theory for typed theories of truth, and of the language of the base theory expanded by the truth predicate for type-free theories. The base theory contains its own (logical, mathematical or empirical[34]) vocabulary, but at the least it has to contain the objects of truth predicate. Using a sufficiently strong mathematical theory as base theory has a great advantage: the truth predicate applies to natural numbers considered as codes of sentences.

Hence, syntactical objects such as sentences are identified with natural numbers. But besides this, truth bearers and the truth predicate itself are syntactical or linguistical entities. And, moreover, they are treated in a syntactical setting. It is exactly this the *adherence* I am talking about. It does not matter if one accepts or avoids ontological commitment towards these objects, namely the objects that can be true. In both case one is inclined to consider the descriptive symbol T and codes of sentences not just a pale theoretical counterpart of the predicate 'to be true' and sentences in

[34] We use Peano arithmetic as the base theory, but other more comprehensive base theories can be used as well.

natural language, precisely because they share their linguistic nature[35]. It is not just as other theories: if we have a first-order formulation of a physical theory there is a formal apparatus with a specific interpretation (arguably more than one) but physical objects — keeping aside the problem of their existence when are not observable — are considered something essentially different from the concepts of the theory. Mathematics is where things get even more complicated, since the intended objects are abstract objects but not yet linguistical entities, so a certain discontinuity does remain. At least one is willing to accept that in a theory of truth what you talk about is not so different in its nature from what you use to talk about. One can object that truth in natural language is a semantic notion and of course it is *also* so. Nevertheless in an axiomatic approach the meaning of truth is given by providing syntactical requirement and it seems that for truth is especially correct that syntax and semantics are very closely related.

This remark forces to reassess the problem of ontological reduction. Since there is a close connection between axiomatic systems and conceptions of truth behind them, a reduction between truth theory cannot be considered something which concerns just theories, it somehow concerns also objects. But it can be seen as an ontological reduction only if you attach to the ontology a very peculiar status, that of set of different linguistic concepts together with laws that regulate their behaviour.

Things change when theories of truth are compared with other kind of theories, in such case the ontological issue assumes a special importance and becomes a problem to be clarified. Whereas reductions of theories of truth to each other can be seen as homogeneous reductions, since usually they share their conceptual apparatus[36], now we are dealing with something different: second-order mathematical theories, i.e. theories of sets of natural numbers, can be reduced to truth theories. How this kind of reduction is carried out? The arithmetical content is left unchanged, just formulae containing second-order parameters are translated into first-order formulae by using truth predicate. The idea is the following: the expression $t \in Y$, such that $\phi(x)$ is the defining formula for the set Y[37], is interpreted as 'the formula Y is true of t', i.e. $T(\ulcorner \phi(t) \urcorner)$, and, respectively $t \notin Y$ is translated with $F(\ulcorner \phi(t) \urcorner)$. Such inhomogeneous translation poses problems on the ontological front. If one accepts Halbach's view[38], ontological reduction in such case consists precisely in the fact that set-theoretical expressions — or assumptions — are turned (by a translation) into truth-theoretical ones. In order to carry out this reduction no other ontological assumptions are

[35] For a defense of the logico-linguistic nature of the notion of truth see Horsten [31], section 5.2.3 p. 91.

[36] A further fine distinction should be made for theories such as RT with more than one truth predicate.

[37] According to the schema of arithmetic comprehension.

[38] See section 2.3 on page 41.

needed, that is, sets are not reduced to some other objects. Once again, ontological reduction concerns ‘words’ more than objects.

Therefore, the field of theories of truth is suitable even to ‘test’ claims about ontological reductions.

Chapter 3

A case study: the proof theory of DT

3.1 Determinate theory of truth

Feferman introduced a new formal theory of truth extending PA in his [18]. As usual, a new unary predicate $T(x)$ for truth is added to the language of PA, where x ranges over codes of sentences in the extended language. The starting point is the informal idea that the domain of the truth predicate consists exactly of the *determinate and meaningful sentences.* In order to formalize this insight a formula $D(x)$, expressing that x is a determinate meaningful sentence, is defined. The system, then, is given by the axioms of PA (in the extended language), axioms for the predicate D and axioms for T relative to D. The resulting system is called DT for *determinate truth.*

3.1.1 Philosophical motivations and background idea

In the first chapter I have outlined different way out from paradoxes in building an axiomatic theory of truth. I have argued that, once a philosophical attitude is chosen, the most natural strategy is to restrict in a suitable way basic principles, that is disquotational sentences. However, things are not simple since such restriction should meet different *desiderata*, namely by restricting the domain of the T-schema one has to:

(i) rule out the liar-like sentences;

(ii) not rule out intuitively unproblematic sentences such as the truth teller or any iteration;

(iii) isolate a natural set of sentences;

(iv) make a philosophically justifiable choice.

The reviewed solutions do not satisfy one or more of this criteria, as TB excludes sentences containing T and it seems to be overmuch restrictive; PTB accepts just T-positive formulae and although this points out an interesting feature of paradoxes (i.e. the fact that the self-reference is harmful just when it is negative) it is hard not to see it as an arbitrary, *ad hoc* choice.

What about DT? The idea is to let the T-schema hold just for sentences in the domain of the truth predicate. This seems very natural, even trivial: each property has a domain, such that it makes sense to apply the property only to objects in that domain. Feferman ascribes the background idea to Russell. In his analysis, Russell blames the reference to a generalized totality of some kind for certain contradictions, among which he mentioned also the Liar paradox. A solution for this problem is to restrict somehow this totality.

> Every propositional function has a certain *range of significance*, within which lie the arguments for which the function has values[1].

It should be observed that Russell refers to *propositional functions* understanding *concepts* in a Fregean sense. In strict logical terms we are talking about *monadic formulae* or, more broadly, about *open formulae*, which denote directly or indirectly properties and relations and among which there is of course $T(x)$. At any rate, the strategy is to reject the assumption that whatever are the objects taken as arguments of a concept, the resulting proposition must be meaningful.

Now, let us investigate the domain of the property 'to be true'. Since the truth predicate is a predicate of some language, arguably, it ranges over the set of sentences of that language. This is not enough to avoid contradiction, so a further restriction must be operated. Feferman identifies a restricted domain with the set of sentences that are *meaningful* and *determinate*, namely those sentences having a definite truth value, true or false[2]. Although, according to Feferman, meaningfulness and determinateness coincide, he keeps them separate because it is a controversial issue whether this equivalence holds. For example, some authors (e.g. Kripke[3]) regard a Liar sentence as meaningful, but of course non determinate (namely it is neither true nor false). The distinction, if there is any, can be roughly formulated as follow: a sentence is meaningful if it "expresses a (linguistically acceptable) proposition" while is determinate or *evaluatability meaningful*[4] if it "expresses a proposition susceptible to receive a truth value". But in order not to commit ourself with consideration of semantical character, this distinction can be

[1] Cfr Russell [41], p. 234.

[2] Reinhardt ([40], p. 220) uses the word *significant* to mean *true or false*. A sentence can be meaningful without being significant, so significant is used in a very restrictive sense, which is exactly the seme of meaningful and determinate in Feferman.

[3] Cfr. [32], ft 17.

[4] For this terminology see McDonald [34], p. 435.

ignored: the sentences in the domain of T will be said both determinate and meaningful.

This is a solution in line with 2^{*5}: critical sentences can be consistently considered as non determinate, they simply are neither true nor false. Accordingly, they cannot be substituted as instances in T-schema. The issue of arbitrariness comes into the picture again, does this choice threat the naturalness of the theory? I shall deal with this question later, in an overall discussion about a critical assessment of the theory.

Let us now see how the domain is formally defined. Let D be the domain of significance of T, and let x be a metavariable that stands for numerals of gödelians of sentences. Then, $D(x)$ can be defined as the disjunction of $T(x)$ and $F(x)$, where $F(x)$ expresses the falsity of x (i.e. the truth of the negation of x). This definition is obtained operatively in the following way: let ϕ be a sentence, since D is the domain of truth, both $T\ulcorner\phi\urcorner \to D\ulcorner\phi\urcorner$ and $F\ulcorner\phi\urcorner \to D\ulcorner\phi\urcorner$ hold. Then

$$T\ulcorner\phi\urcorner \vee F\ulcorner\phi\urcorner \to D\ulcorner\phi\urcorner.$$

Conversely, the T-schema is restricted to sentences satisfying D: $D\ulcorner\phi\urcorner \to (T\ulcorner\phi\urcorner \leftrightarrow \phi)$ and, obviously, $D\ulcorner\phi\urcorner \to (F\ulcorner\phi\urcorner \leftrightarrow \neg\phi)$. Thus, we have: $D\ulcorner\phi\urcorner \to (T\ulcorner\phi\urcorner \vee F\ulcorner\phi\urcorner \leftrightarrow \phi \vee \neg\phi)$. Then:

$$D\ulcorner\phi\urcorner \to T\ulcorner\phi\urcorner \vee F\ulcorner\phi\urcorner.$$

The conclusion is:

$$D\ulcorner\phi\urcorner \leftrightarrow T\ulcorner\phi\urcorner \vee F\ulcorner\phi\urcorner \quad \text{for each sentence } \phi.$$

This argument allows us to identify D in terms of T: $D(x)$ just abbreviates $T(x) \vee F(x)$ and so it must not be introduced separately as a predicate[6].

Moreover, Feferman requires that $D(x)$ must be *strongly compositional*, in a way that will now be explained. On the one hand, D must be closed under the propositional operations (e.g. if a sentence belongs to D, then its negation belongs to it as well) and quantifiers. For the universal quantifier, we take a compound sentence to belong to D if all its substitution instances by meaningful terms belong to D. On the other hand, the closure conditions for D are assumed to be invertible: a complex sentence is meaningful and determinate *only if* its syntactic constituent sentences are so. This is not a trivial claim. Of course it holds for meaningfulness: a sentence is meaningful only if all of its parts are meaningful. That is, if a negation of

[5] See section 1.4 on page 14.

[6] Fujimoto [22] ft 22, pointed out that $D\ulcorner\phi\urcorner \leftrightarrow T\ulcorner\phi\urcorner \vee F\ulcorner\phi\urcorner$ for each (standard) sentence ϕ does not entail $D(x) \leftrightarrow T(x) \vee F(x)$ for all x because of the overspill argument. Being a theory extending PA, DT must have a non-standard model and thus non-standard sentences for which the equivalence is not necessarily provable.

a sentence is meaningful then so is the sentence itself. What about determinateness? By the common interpretation of connectives, a sentence like $\phi \lor \psi$ is determinate (e.g. true) even when one of the constituents ϕ or ψ has not a truth value, while the other is true. Nevertheless, Feferman opts for a strictly compositional semantics: irrespective of the logical form of a compound sentence, the presence of just an undeterminate constituent it is enough to make the whole sentence undeterminate too. According to this requirement, a suitable logic is chosen to deal with the lack of truth value so that a sentence is evaluated as neither true nor false if one of its components lacks a truth value; that is a logic like the Weak Kleene one (WKL). The strong compositionality of D is expressed by formulae with a biconditional like:

$$D(x \underset{\cdot}{\lor} y) \leftrightarrow D(x) \land D(x).$$

Anyhow, we shall see that the requirement of strong compositionality it is not met in full because of the treatment of $\rightarrow$.

3.1.2 Axiomatization

Now we can turn the preceding intuitive ideas into axioms in a formalized setting and later sound out how close we are to them. Let $\mathcal{L}_{\mathsf{PA}}$ be the language of PA and let $\mathcal{L}_T = \mathcal{L}_{\mathsf{PA}} \cup \{T\}$ be the extension of $\mathcal{L}_{\mathsf{PA}}$ by a new unary predicate T. Moreover, let $\mathrm{Sent}_{\mathcal{L}}$ be the representation of the syntactical notion that holds of x if and only if x is a code of a sentence of $\mathcal{L}$. As usual we can abbreviate $T(\underset{\cdot}{\neg} x)$ by using $F(x)$.

Definition 3.1.1. The system DT consists of all axioms of PAT (i.e. PA plus full induction for formulae of $\mathcal{L}_T$) and the following axioms:

(DT1) $\forall x\ [\mathrm{AtSent}_{\mathsf{PA}}(x) \rightarrow D(x)]$

(DT2) $\forall t\ [D(\underset{\cdot}{T}(t)) \leftrightarrow D(t^{\circ})]$

(DT3) $\forall x\ [\mathrm{Sent}_T(x) \rightarrow (D(\underset{\cdot}{\neg} x) \leftrightarrow D(x))]$

(DT4) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\lor} y) \rightarrow (D(x \underset{\cdot}{\lor} y) \leftrightarrow D(x) \land D(y))]$

(DT5) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\rightarrow} y) \rightarrow (D(x \underset{\cdot}{\rightarrow} y) \leftrightarrow D(x) \land (T(x) \rightarrow D(y)))]$

(DT6) $\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\forall} v x) \rightarrow (D(\underset{\cdot}{\forall} v x) \leftrightarrow \forall t D(x[t/v]))]$

(DT7) $\forall s \forall t\ [T\ulcorner s \underset{\cdot}{=} t \urcorner \leftrightarrow s^{\circ} = t^{\circ}]$

(DT8) $\forall t\ [D(t^{\circ}) \rightarrow (T(\underset{\cdot}{T} t) \leftrightarrow T(t^{\circ}))]$

(DT9) $\forall x\ [\mathrm{Sent}_T(x) \land D(x) \rightarrow (T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T(x))]$

(DT10) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\lor} y) \land D(x \underset{\cdot}{\lor} y) \rightarrow (T(x \underset{\cdot}{\lor} y) \leftrightarrow T(x) \lor T(y))]$

(DT11)$\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\to} y) \wedge D(x \underset{\cdot}{\to} y) \to (T(x \underset{\cdot}{\to} y) \leftrightarrow T(x) \to T(y))]$

(DT12)$\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\forall} vx) \wedge D(\underset{\cdot}{\forall} vx) \to (T(\underset{\cdot}{\forall} vx) \leftrightarrow \forall t T(x[t/v]))]$

The first group concerns the predicate D: (DT1) states that every atomic sentence of the base language satisfies $D(x)$, but it can be easily shown by induction on the complexity of ϕ that DT proves $D\ulcorner\phi\urcorner$ for each sentence ϕ of $\mathcal{L}_{\mathsf{PA}}$. The second axiom says that a sentence of the form 'something is true' is determinate if and if what is said true is already determinate. The remaining axioms reflect the closure conditions for D: it satisfies the strongly compositional conditions. Nevertheless, axiom (DT5) seems to express an anomaly with respect to the claim of strict compositionality, and indeed so it is as we shall see in the next section. The second group expresses the compositional nature of truth in a quite restricted manner: for x satisfying D, $T(x)$ satisfies the usual recursive defining conditions (Tarski's clauses). Moreover, under the same restriction, truth is materially adequate in a Tarskian sense: DT proves

$$D\ulcorner\phi\urcorner \to (T\ulcorner\phi\urcorner \leftrightarrow \phi) \text{ for each } \phi \text{ of } \mathcal{L}_T.$$

This can be shown in DT, again, by induction on the formation of ϕ.

3.1.3 Critical assessment

The system DT is type-free, i.e. it contains its own truth predicate that can be applied also to sentences which already contain it. Moreover, the chosen base logic for the system is ordinary classical 2-valued logic. Inconsistency is avoided by restricting T-schema as follows:

$$D\ulcorner\phi\urcorner \to (T\ulcorner\phi\urcorner \leftrightarrow \phi) \text{ for each } \phi \text{ of } \mathcal{L}_T.$$

Of course, it can be shown that the Liar sentence (in its different formulations) does not fall within the scope of D, that is it must be neither true nor false.

Let λ be a sentence such that $\mathsf{DT} \vdash \lambda \leftrightarrow F\ulcorner\lambda\urcorner$. If $D\ulcorner\lambda\urcorner$, then the usual contradiction from the T-schema is derived. So DT proves $\neg D\ulcorner\lambda\urcorner$ and the paradox cannot be carried out. Note that the same holds when λ is such that $\mathsf{DT} \vdash \lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$.

It is worth noting how the strengthened liar is accounted for in DT. There are different formulations of this sentence, consider the following:

This sentence is not true, that is, it is false or meaningless.

In formal terms, let σ be a sentence of $\mathcal{L}_T$ such that $\mathsf{DT} \vdash \sigma \leftrightarrow \neg D\ulcorner\sigma\urcorner \vee F\ulcorner\sigma\urcorner$. Again the escape route is to prove that $\neg D\ulcorner\sigma\urcorner$. Suppose $D\ulcorner\sigma\urcorner$, then $\mathsf{DT} \vdash \sigma \leftrightarrow F\ulcorner\sigma\urcorner$, and we have the classical contradiction as before. Hence $\neg D\ulcorner\sigma\urcorner$. Then $\mathsf{DT} \vdash \sigma$ but $\mathsf{DT} \nvdash D\ulcorner\sigma\urcorner$.

In a very elegant manner paradoxes are avoided, moreover DT can be proved consistent by providing a standard model for it[7]. Anyhow, we saw that there are other different desirable criteria, As a reminder we repeat them, in Leitgeb's formulation:

(a) Truth should be expressed by a predicate (and a theory of syntax should be available).

(b) If a theory of truth is added to mathematical or empirical theories, it should be possible to prove the latter true.

(c) The truth predicate should not be subject to any type restrictions.

(d) T-biconditionals should be derivable unrestrictedly.

(e) Truth should be compositional.

(f) The theory should allow for standard interpretations.

(g) The outer logic and the inner logic should coincide.

(h) The outer logic should be classical.

Let us now how close DT goes to meeting them:

(a) Met.

(b) Met for PA.
This requirement is somehow related to the formulation of axiom (DT5). We should be able to prove in DT that each theorem of PA is true, that is:

$$\mathsf{DT} \vdash T(\forall x\ (\mathrm{Sent}_{\mathsf{PA}}(x) \wedge \mathrm{Bew}_{PA}(x) \rightarrow T(x))), \tag{3.1}$$

where $\mathrm{Sent}_{\mathsf{PA}}(x)$ expresses that x is the Gödel number of a sentence of $\mathcal{L}_{\mathsf{PA}}$ and $\mathrm{Bew}(x)$ is the primitive recursive representation of provability predicate for PA. If we admitted the full interdefinability of connectives, there would be a problem about that.

Reasoning informally, the claim (3.1) can be proved by induction in DT. But in order to let the truth predicate commute with the universal quantifier we need to apply the axiom (DT12). To this aim, it must be proved that:

$$D(\forall x\ (\mathrm{Sent}_{\mathsf{PA}}(x) \wedge \mathrm{Bew}_{PA}(x) \rightarrow T(x))),$$

or equivalently modulo the axiom (DT6):

$$D(\mathrm{Sent}_{\mathsf{PA}}(\bar{n}) \wedge \mathrm{Bew}_{PA}(\bar{n}) \rightarrow T(\bar{n})) \quad \text{for each numeral } \bar{n}.$$

[7] See the next section.

If $\to$ is defined as usual in terms of $\neg$ and $\vee$, then modulo logical equivalence we have to prove for each numeral $\bar{n}$:

$$D(\neg(\mathrm{Sent}_{\mathsf{PA}}(\bar{n}) \wedge \mathrm{Bew}_{PA}(\bar{n})) \vee T(\bar{n})).$$

But this disjunction is determinate if and only if both members are so, this requires that $D(T(\bar{n}))$ must hold for each $\bar{n}$, where $\bar{n}$ is a whatever numeral, no longer the code of a sentence provable in PA. So, in particular, we must have $D(T\ulcorner\lambda\urcorner)$, where λ is the liar sentence, that is a sentence for which DT proves $\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$. This leads to contradiction because if $D(T\ulcorner\lambda\urcorner)$ holds, then by (DT2) $D\ulcorner\lambda\urcorner$ holds as well. And so we would get $\lambda \leftrightarrow T\ulcorner\lambda\urcorner$. In order to avoid this problem, Feferman takes $\to$ as a separate propositional operation, so that $D(x \underset{\cdot}{\rightarrow} y) \not\leftrightarrow D(\underset{\cdot}{\neg} x \underset{\cdot}{\vee} y)$. This leads to the formulation of the axiom (DT5) in which the predicate D applied to conditionals is not simply equivalent to the conditional with D distributed over members: the determinateness of the consequent holds under the hypothesis that the antecedent is not just determinate but true. In this case we get $D(T(\bar{n}))$ only when $\bar{n}$ is the code of a sentence of the base language that is provable in PA and this is not the case for λ. Accordingly, this formulation is not fully compositional. At any rate, the logic of $\to$ is left unchanged and although D does not meet the strong compositionality with respect to $\to$, T does as the axiom (DT11) states.

(c) Met.

(d) Met only for those sentences satisfying the D predicate.

(e) Met under the same restriction.

(f) Met.

(g) Met only to the extent that the inner logic is classical for sentences satisfying the D predicate.
DT is formulated in classical logic, at the same time there are sentences which lack to receive a truth value, i.e. $\exists x(\mathrm{Sent}_T(x) \wedge \neg D(x))$. So, in some way the truth predicate is partial (it is better to say that T is total on a restricted domain): the inner logic is a three-valued one. Apparently, it seems that DT shows the same asymmetry of KF, with the proviso that Feferman uses a different logic to handle with trivalence (a variant of WKL with a different interpretation for the conditional.). Nevertheless, it is clear from the axioms of compositionality for T that for those sentences satisfying D the inner logic is classical. For example $\mathsf{DT} \vdash \lambda \vee \neg\lambda$ and, since $\mathsf{DT} \vdash \neg D\ulcorner\lambda\urcorner$, $\mathsf{DT} \vdash \neg T\ulcorner\lambda \vee \neg\lambda\urcorner$. Therefore, working internally to D, the condition of bivalence is trivially restored.

(h) Met.

Once one accepts the justifiability of the underlying choice of DT, the restrictions in (d) and (e) becomes very natural as all the principles characterizing the truth of a sentence (disquotation, compositionality etc.) must be subordinated to the claim about the determinateness of that sentence, in other words “the conditions on D should be prior to those on T, that is, determinate meaningfulness is prior to truth[8]”. For example we have $D(x\underset{\cdot}{\vee}y) \leftrightarrow D(x) \wedge D(y)$ and then $D(x\underset{\cdot}{\vee}y) \rightarrow (T(x\underset{\cdot}{\vee}y) \leftrightarrow T(x) \vee T(y))$.

Hence, the issue of naturalness for restrictions to (d) and (e) is subsumed by the issue of naturalness for the whole theory DT. To restrict the domain of T is a choice that can be further questioned by using natural language as touchstone. It can be argued that whatever (name of) sentence could be said true (or not true) and that there are no restrictions of any kind for the domain. But saying something true is quite different from saying something *meaningfully* true. Does it make sense to say the liar sentence true or false? Natural languages are characterized by freedom in predications, even if the resulting sentence is meaningless (this is also their beauty). Formal setting offers us a further chance: to totally exclude the possibilities of meaningless predications, as far as what is acceptable has to be specified before.

It seems less simple to justify the fact that DT does not meet (g), but it can be simply argued that DT works in a desirable way just internally to the logic of D.

3.1.4 Consistency

The proof of the consistency of DT is given by providing a standard model for it. The starting point is a standard model $\mathbb{N}$ for PA which has to be expanded to a model $\mathcal{M}$ for DT in the language $\mathcal{L}_T$. In this construction, the intermediate step will be a 3-valued model $\mathcal{M}^*$ from which $\mathcal{M}$ will be obtained. $\mathcal{M}^*$ will be given by an assignment $v\colon \mathrm{Sent}_T \rightarrow \mathbf{3}$, where $\mathbf{3} = \{t, f, u\}$. I specify now a distinction that will be ignored in the notation:

- $\underline{\neg}$, $\underline{\vee}$, $\underline{\rightarrow}$, $\prod$: operations on $\mathbf{3}$ between truth-values. These are evaluated by using a variant of weak Kleene semantics, as we shall see in Definition 3.1.2.
- $\neg$, $\vee$, $\rightarrow$, $\forall$: operations between sentences of $\mathcal{L}_T$. In evaluating resulting compound sentences we shall refer to the corresponding operations on $\mathbf{3}$.

Since there is no risk of ambiguity, I drop the bar under propositional operations and write simply $\neg$, $\vee$, $\rightarrow$ even for operations on $\mathbf{3}$.

Definition 3.1.2. Let $D(a)$ be $a = t$ or $a = f$ for $a \in \mathbf{3}$.

[8] Cfr. Feferman [18], p.207.

(i) $D(\neg a)$ iff $D(a)$;

	$\neg a$
t	f
f	t
u	u

(ii) $D(a \vee b)$ iff $D(a)$ & $D(b)$;

$a \vee b$	t	f	u
t	t	t	u
f	t	f	u
u	u	u	u

(iii) $D(a \to b)$ iff $D(a)$ & $(a = f$ or $D(b))$;

$a \to b$	t	f	u
t	t	f	u
f	t	t	**t**
u	u	u	u

(iv) $D(\prod \{ a_i \mid i \in I \})$ iff $D(a_i)$ for each $i \in I$; if $D(\prod \{ a_i \mid i \in I \})$, then $\prod \{ a_i \mid i \in I \} = t$ iff for each $i \in I$, $a_i = t$, else $\prod \{ a_i \mid i \in I \} = u$.

Define, then, $a \wedge b$ and $\sum \{ a_i \mid i \in I \}$ as usual in terms of $\neg$, $\vee$ and $\forall$. We cannot eliminate $\to$ in favor of $\neg$, $\vee$ as in the weak Kleene semantics: $a \to b$ is determined by a truth table which differs from that of $\neg a \vee b$ for the value in bold.

The 2-valued model $\mathbb{N}$ of PA is expanded to a 3-valued model by using a Krikpe-style construction for a nonhierarchical truth predicate[9]. In Kripke's semantics T is the only partially defined predicate of the language $\mathcal{L}_T = \mathcal{L}_{PA} \cup \{T\}$. The interpretation of the truth predicate is given by a pair (S_1, S_2), where S_1 and S_2 are, respectively, the extension and the antiextension of the truth predicate. The condition that makes T a partial predicate is the following: $S_1 \cup S_2$ does not exhaust the domain — the set of sentences. We start building a (partial) model for $\mathcal{L}_T$: since the arithmetical vocabulary is interpreted in the standard way, we can expand a model of PA by adding the interpretation of the truth predicate. Thus, a partial model will be a triple $(\mathbb{N}, S_1, S_2)$, in which we have just to define suitable extensions and antiextensions.

The Kripke's construction proceeds inductively. At stage 0 we assume to be in a state of maximum ignorance: both the extension and the antiextension of T are empty, i.e.

$$\mathcal{M}_0 = (\mathbb{N}, \emptyset, \emptyset).$$

Suppose we have defined $\mathcal{M}_\alpha = (\mathbb{N}, S_1, S_2)$ for any α, then $\mathcal{M}_{\alpha+1}$ is the triple $(\mathbb{N}, S_1^{'}, S_2^{'})$, where

$$S_1^{'} = \{ \phi \in \mathcal{L}_T \mid \mathcal{M}_\alpha \models \phi \}$$

[9] Cfr. Kripke [32].

and

$$S_2^{'} = \{\, \phi \in \omega \mid \phi \notin \mathrm{Sent}_T \,\} \cup \{\, \phi \in \mathcal{L}_T \mid \mathcal{M}_\alpha \models \neg\phi \,\}\,.$$

That is, the extension of T at the stage α is the set of (codes of) sentence that are true in the model of the previous stage, and the antiextension is the set of elements of the domain which either are not (codes of) sentences or are (codes of) sentences that are false in $\mathcal{M}_\alpha$. The definition of satisfaction is given by using an appropriate evaluation schema, the strong Kleene schema, which allows us to deal with sentences that might fail to receive a classical truth value. By going on in this construction, a sequence of models is built and more and more sentences of $\mathcal{L}_T$ are added to the extension or to the antiextension of T. We can define a model $\mathcal{M}_\alpha$ for each ordinal α. A transfinite chain is built, by taking the union of the previous extensions and antiextensions at limit stages, that is, for ordinals that are not successor ordinals.

Does this process eventually come to the end? The affermative answer is closely related with an important monotonicity property of the Strong Kleene evaluation schema. Let Φ be an operator on pairs of sets of natural numbers defined in the following way:

$$\Phi((S_1, S_2)) = (S_1^{'}, S_2^{'}),$$

and let $\leq$ be a partial ordering on these pairs such that:

$$(S_1, S_2) \leq (S_1^{\circ}, S_2^{\circ}) \text{ iff } S_1 \subseteq S_1^{\circ} \text{ and } S_2 \subseteq S_2^{\circ}.$$

It can be proved that Φ is a monotone (order-preserving) operation on $\leq$, that is: if $(S_1, S_2) \leq (S_1^{\circ}, S_2^{\circ})$ then $\Phi((S_1, S_2)) \leq \Phi((S_1^{\circ}, S_2^{\circ}))$. This means that when we extend the interpretation of T at each stage, all truth values previously established do not change, at most certain undefined truth values become defined. In other words, as α increases, the predicate T increases in both its extension and its antiextension. The set of sentences is like a container that is progressively emptied at each level: this leads to an ordinal for which no new sentences can be declared true or false, namely there should be an ordinal level α for which:

$$(S_{1\alpha}, S_{2\alpha}) = (S_{1\alpha+1}, S_{2\alpha+1}).$$

This pair is a fixed point of the operator Φ, and the matching model is called *fixed point model* of $\mathcal{L}_T$. The existence of this model can be proved formally, and, moreover, it can also be proved that it is a "minimal" fixed point, in the sense that any fixed point extends it.

Feferman takes the least fixed point $\mathcal{M}^* = (\mathbb{N}, S_1^*, S_2^*)$ as a 3-valued model of $\mathcal{L}_T$. Note that Kripke uses strong Kleene semantics in his construction as evaluation schema, while Feferman uses a variant of the weak Kleene semantics. How can we ensure that changing our evaluation rules a minimum fixed point nevertheless exists? Kripke pointed out that the construction can by carried out as well as far as the monotonicity of the operator Φ is preserved:

> So far we have assumed that truth gaps are to handled according to the methods of Kleene. It is by no means necessary to do so. Just about any schema for handling truth-value gaps is usable, provided that the basic property of the monotonicity of Φ is preserved; that is, provided that extending the interpretation of $T(x)$ never changes the truth value of any sentence of $\mathcal{L}$, but at most gives truth values to previously undefined cases. Given any such schema, we can use the previous arguments to construct the minimal fixed point and other fixed points, define the levels of sentences and the notions of 'grounded', 'paradoxical', etc.[10]

Now, I want to briefly explain how the monotonicity of the defined operations $\neg$, $\vee$, $\rightarrow$ and Π ensures the monotonicity of Φ. Let $(S_{1\alpha}, S_{2\alpha})$ and $(S_{1\beta}, S_{2\beta})$ be two pairs such that the latter extends the former, i.e.:

$$S_{1\alpha} \subseteq S_{1\beta} \text{ and } S_{2\alpha} \subseteq S_{2\beta}.$$

Now, the monotonicity of Φ, i.e.

$$\Phi((S_{1\alpha}, S_{2\alpha})) \leq \Phi((S_{1\beta}, S_{2\beta})),$$

entails:

$$(S^{'}_{1\alpha}, S^{'}_{2\alpha}) \leq (S^{'}_{1\beta}, S^{'}_{2\beta}),$$

that is

$$S^{'}_{1\alpha} \subseteq S^{'}_{1\beta} \text{ and } S^{'}_{2\alpha} \subseteq S^{'}_{2\beta}.$$

Let us focus on the extension (but the same holds for the antiextension): $S^{'}_{1\alpha} \subseteq S^{'}_{1\beta}$ means that

$$\{ \phi \mid \mathcal{M}_\alpha = (\mathbb{N}, S_{1\alpha}, S_{2\alpha}) \models \phi \} \subseteq \{ \phi \mid \mathcal{M}_\beta \models \phi \}.$$

In other words, the condition of monotonicity of Φ expresses that any sentence true (or false) in $\mathcal{M}_\alpha$ retains its truth value in $\mathcal{M}_\beta$. At this point, evaluations come into the picture: let v_α and v_β be the evaluations corresponding to the models $\mathcal{M}_\alpha$ and $\mathcal{M}_\beta$, respectively. The condition $S^{'}_{1\alpha} \subseteq S^{'}_{1\beta}$ and its analogous for the antiextension can be read as: the formulae evaluated by v_β must be all those evaluated by v_α (with the same truth value) plus a number of formulae that were undefined for v_α and that become true or false in v_β. Since for all ϕ, $v_\alpha(\phi)$ and $v_\beta(\phi) \in \{t, f, u\}$ we can express the inclusion between sets of formulae above by saying:

$$v_\alpha(\phi) \leq v_\beta(\phi) \text{ for any sentence } \phi,$$

[10] Cfr. [32], p. 711.

where $\leq$ is the partial reflexive ordering of the set **3**. The reflexivity condition (i.e. $u \leq u$, $t \leq t$ and $f \leq f$) ensures that the evaluated sentences retain their truth values and the other conditions ($u \leq t$, $u \leq f$) ensure that no truth value established by v becomes undefined; at most certain previously undefined truth values become defined. And this is exactly the property of monotonicity of Φ. Since ϕ can be a compound sentence, in order to state the monotonicity of Φ the operations between truth values (whatever semantics is chosen) must preserve the ordering of **3**.

It can be proved that this ordering is preserved by the Feferman operations:

Lemma 3.1.1. Each of $\neg$, $\vee$, $\rightarrow$ and Π is monotonic on the reflexive ordering of $\{t, f, u\}$.

Proof. I just focus on the monotonicity of $\rightarrow$, the non trivial case. Assume that $a \leq a'$ and $b \leq b'$ for any a, a', b, b' in **3**. In order to show that:

$$(a \rightarrow b) \leq (a' \rightarrow b'),$$

we distinguish cases on the possible values of a and b

- $D(a)$ and $D(b)$.
  If both a and b have a determinate truth value, then $a = a'$ and $b = b'$, thus trivially $(a \rightarrow b) = (a' \rightarrow b')$.

- $a = u$.
  From the definition we have, whatever b is, $(u \rightarrow b) = u$ and u is $\leq$ any value.

- $D(a)$ and $b = u$.
  If $a = t$, then $(t \rightarrow u) = u$, that again is $\leq$ any value. If $a = f$, then $(f \rightarrow u) = t$ and $(f \rightarrow b') = t$ whatever b'.

□

Based on what we said, this lemma yields an expansion of $\mathbb{N}$ to a fixed point model $\mathcal{M}^* = (\mathbb{N}, S_1^*, S_2^*)$. In it each sentence ϕ is evaluated according to the rules given in definition 3.1.2 and, furthermore, the value of $T\ulcorner\phi\urcorner$ is the same as ϕ.

Theorem 3.1.1. There is a 3-valued model $\mathcal{M}^*$ of $\mathcal{L}_T$ given by an assignment $v(\phi)$ in $\{t, f, u\}$ to each sentence ϕ of $\mathcal{L}_T$ satisfying the following conditions:

(i) (a) If $\phi \in \text{AtSent}_{\mathcal{L}}$, then $v(\phi) = t$ or $v(\phi) = f$ and $v(\phi) = t$ iff $\mathbb{N} \models \phi$.

(b) If $\phi \in \text{Sent}_{\mathcal{L}_T}$ and ϕ is denoted by the term $\ulcorner\phi\urcorner$, then $v(T\ulcorner\phi\urcorner) = v(\phi)$ otherwise $v(T\ulcorner\phi\urcorner) = f$.

(ii) $v(\neg\phi) = \neg v(\phi)$.

(iii) $v(\phi \vee \psi) = v(\phi) \vee v(\psi)$.

(iv) $v(\phi \to \psi) = v(\phi) \to v(\psi)$.

(v) $v(\forall x \chi(x)) = \prod \{\, v(\chi(n)) \mid n \in N \,\}$.

$\mathcal{M}^*$ is then converted into a 2-valued model $\mathcal{M} = (\mathbb{N}, T)$ of $\mathcal{L}$. Intuitively the approach is the following: T is the only partially interpreted symbol, but it is enough to make $\mathcal{M}^*$ a partial model. Since classical models are characterized by the fact that the antiextension of a property coincides with the complement of the extension with respect to the domain, $\mathcal{M}^*$ is converted into a classical model $\mathcal{M}$ by putting in the extension of T all and only the (codes of) sentences true in $\mathcal{M}^*$, i.e. $\mathcal{M} \models T\ulcorner\phi\urcorner$ if and only if $v(T\ulcorner\phi\urcorner) = t$ in $\mathcal{M}^*$. By doing so, we build a model in which T is a total predicate. Satisfaction in $\mathcal{M}$ is specified as follows:

Definition 3.1.3. (i) If $\phi \in \mathrm{AtSent}_T$, then $\mathcal{M} \models \phi$ iff $v(\phi) = t$ in $\mathcal{M}^*$.

(ii) $\mathcal{M} \models \neg\phi$ iff not $\mathcal{M} \models \phi$.

(iii) $\mathcal{M} \models \phi \vee \psi$ iff $\mathcal{M} \models \phi$ or $\mathcal{M} \models \psi$.

(iv) $\mathcal{M} \models \phi \to \psi$ iff not $\mathcal{M} \models \phi$ or $\mathcal{M} \models \psi$.

(v) $\mathcal{M} \models \forall x \chi(x)$ iff $\mathcal{M} \models \chi(\bar{n})$ for each $n \in N$.

In order to state the consistency of DT, we have to show that $\mathcal{M}$ is a model for it. That is, the axioms of DT must be all true in $\mathcal{M}$.

Lemma 3.1.2. (i) $\mathcal{M} \models T\ulcorner\phi\urcorner$ iff $v(\phi) = t$.

(ii) $\mathcal{M} \models F\ulcorner\phi\urcorner$ iff $v(\phi) = f$.

(iii) $\mathcal{M} \models D\ulcorner\phi\urcorner$ iff $D(v(\phi))$.

(iv) $\mathcal{M} \models D\ulcorner T\ulcorner\phi\urcorner\urcorner$ iff $D(v(\phi))$.

Proof. (i) $\mathcal{M} \models T\ulcorner\phi\urcorner \Leftrightarrow v(T\ulcorner\phi\urcorner) = t$ in $\mathcal{M}^* \Leftrightarrow v(\phi) = t$.

(ii) Analogous modulo condition (ii) of the theorem 3.1.1 and (DT9).

(iii) It easily follows from (i) and (ii) of this lemma, (iii) of definition 3.1.3 and condition (iii) of the theorem 3.1.1.

(iv) $\mathcal{M} \models D\ulcorner T\ulcorner\phi\urcorner\urcorner \Leftrightarrow \mathcal{M} \models T\ulcorner T\ulcorner\phi\urcorner\urcorner$ or $\mathcal{M} \models F\ulcorner T\ulcorner\phi\urcorner\urcorner \Leftrightarrow v(T\ulcorner\phi\urcorner) = t \vee v(T\ulcorner\phi\urcorner) = f \Leftrightarrow v(\phi) = t \vee v(\phi) = f := D(v(\phi))$.

□

Theorem 3.1.2. $\mathcal{M}$ is a model of DT.

Proof. We prove that the axioms of DT are true in $\mathcal{M}$.

(DT1) For any atomic sentence ϕ of $\mathcal{L}$: $\mathcal{M} \models D\ulcorner\phi\urcorner$.
If ϕ is an atomic sentence of the base language, for (a) of 3.1.1: $v(\phi) = t$ or $v(\phi) = f$ in $\mathcal{M}^*$, that is $D(v(\phi))$. For condition (iii) of lemma 3.1.2: $\mathcal{M} \models D\ulcorner\phi\urcorner$.

(DT2) For any term s of $\mathcal{L}_T$: $\mathcal{M} \models D(T(s)) \leftrightarrow D(s^\circ)$.
$\mathcal{M} \models D\ulcorner T(s)\urcorner$ entails $\mathcal{M} \models T\ulcorner T(s)\urcorner$ or $\mathcal{M} \models F\ulcorner T\ulcorner s\urcorner\urcorner$. This holds if and only if $v(T(s)) = t$ or $v(T(s)) = f$. Let s° be the value of s, then for (b) $v(s^\circ) = t$ or $v(s^\circ) = f$, thus $D(v(s^\circ))$, which yields $\mathcal{M} \models D(s^\circ)$.

(DT3) – **(DT6)** I shall show just **(DT5)**: $\mathcal{M} \models \forall x, y\ [D(x \dot{\vee} y) \leftrightarrow D(x) \wedge D(y)]$.
Let ϕ and ψ sentences of $\mathcal{L}_T$:
$\mathcal{M} \models D\ulcorner\phi\dot{\vee}\psi\urcorner \Leftrightarrow D(v(\phi\dot{\vee}\psi)) \Leftrightarrow D(v(\phi) \vee v(\psi)) \underset{Def 3.1.2}{\Longleftrightarrow} D(v(\phi))\ \&\ D(v(\psi)) \Leftrightarrow \mathcal{M} \models D\ulcorner\phi\urcorner\ \&\ \mathcal{M} \models D\ulcorner\psi\urcorner \Leftrightarrow \mathcal{M} \models D\ulcorner\phi\urcorner \wedge D\ulcorner\psi\urcorner$.

(DT7) For any t and s such that $D\ulcorner t = s\urcorner$ $\mathcal{M} \models T\ulcorner t = s\urcorner \leftrightarrow t^\circ = s^\circ$.
$\mathcal{M} \models T\ulcorner t = s\urcorner \Leftrightarrow v(t = s) = t$ in $\mathcal{M}^*$. Since $t = s$ is an atomic sentence of $\mathcal{L}$ from (a) we have $\mathbb{N} \models t = s$. Thus $t^\circ = s^\circ$.

(DT8) For any term t such that $D(t)$: $\mathcal{M} \models T\ulcorner T(t)\urcorner \leftrightarrow T(t^\circ)$.
$\mathcal{M} \models T\ulcorner T(t)\urcorner \Leftrightarrow v(T(t)) = t$ in $\mathcal{M}^* \Leftrightarrow v(t^\circ) = t \Leftrightarrow \mathcal{M} \models T(t^\circ)$.

(DT9) – **(DT12)** Just an example: **(DT10)**. For any sentence ϕ, ψ of $\mathcal{L}_T$: $\mathcal{M} \models D\ulcorner\phi\dot{\vee}\psi\urcorner \to (T\ulcorner\phi\dot{\vee}\psi\urcorner \leftrightarrow T\ulcorner\phi\urcorner \vee T\ulcorner\psi\urcorner)$.
Assume $D\ulcorner\phi\dot{\vee}\psi\urcorner$ and $T\ulcorner\phi\dot{\vee}\psi\urcorner$. $\mathcal{M} \models T\ulcorner\phi\dot{\vee}\psi\urcorner \Leftrightarrow v(\phi\dot{\vee}\psi) = t$ in $\mathcal{M}^* \Leftrightarrow v(\phi) \vee v(\psi) = t$. The hypothesis $\mathcal{M} \models D\ulcorner\phi\dot{\vee}\psi\urcorner$ with **(DT5)** yields: $\mathcal{M} \models D\ulcorner\phi\urcorner \wedge D\ulcorner\psi\urcorner \Leftrightarrow \mathcal{M} \models D\ulcorner\phi\urcorner$ and $\mathcal{M} \models D\ulcorner\psi\urcorner \Leftrightarrow D(v(\phi))$ and $D(v(\psi)) \Leftrightarrow D(v(\phi) \vee v(\psi))$. This condition allows us to use the definition 3.1.2, and obtain: $v(\phi) \vee v(\psi) = t \Leftrightarrow v(\phi) = t$ or $v(\psi) = t \Leftrightarrow \mathcal{M} \models T\ulcorner\phi\urcorner$ or $\mathcal{M} \models T\ulcorner\psi\urcorner \Leftrightarrow \mathcal{M} \models T\ulcorner\phi\urcorner \vee T\ulcorner\psi\urcorner$. □

Thus, we have proved by providing a (standard) model that DT is consistent.

3.2 Proof theory of the determinate theory of truth

In the second chapter I have argued in favour of the relevance of a metatheoretical inquiry both from a philosophical and instrumental point of view. Let us now outline an example of how such analysis is carried out, with respect to the just described theory DT. First we shall see the position of DT in the galaxy of axiomatic theories of truth introduced before and, then, we shall investigate proof-theoretic power of DT by reducing it to a particular mathematical system. I shall discuss the value of the achieved results of reduction.

3.2.1 Relating DT to other theories of truth

Theories of truth can be compared to each other by means of different notions, the most widely used are those presented previously. The tool we are going to use throughout our discussion is *relative truth-definability*[11]. This choice is justified by different reasons:

- It is a refined instrument which is capable of establishing strong relations between theories by distinguishing theories that are joined by other notions of reduction.
- Being stronger, it implies other notions of reducibility:
  - Truth-definability implies conservativity on $\mathcal{L}_{\mathsf{PA}}$: if a theory S is relatively truth-definable in T then the truth-free theorems of the former are included in the latter, i.e. S is conservative over T for formulae in $\mathcal{L}_{\mathsf{PA}}$. This holds since theories share their base theory and relative truth-definability leaves the arithmetical content unchanged[12].
  - A relative truth-definition is a (strict) relative interpretation, or better an interpretation without relativization of quantifiers. So, if the truth predicate of a theory S is definable in T then S is relatively interpretable in T.
  - This does not hold in general for proof-theoretic reducibility, but as far as truth theories are concerned this is the case[13].
- It is particularly suitable for a philosophical approach to metatheoretical investigation. Other notions of reduction focus on the truth-free content of theories, but theories sharing the same arithmetical content can hide very different ways of understanding truth. Since in a truth theory the underlying conception of truth is embodied into basic principles, it is attractive to investigate the ability of a theory to simulate the way in which the other works. To this extent relative truth-definability is an helpful tool in comparing conceptual aspects of truth theories.

In the following diagram we can see the position of DT (and $\mathsf{DT}\restriction$) with respect to the other introduced theories.
In the diagram $\mathsf{S} \Rightarrow \mathsf{T}$ means that S is truth definable in T and the reverse direction does not hold. And $\mathsf{S} \Rrightarrow \mathsf{T}$ means that it is still open whether

[11] See definition 2.4.1 on page 49.

[12] However, this holds in general when the base theory B of S and B' of T are different. In this case S can be truth-definable in T only under the condition that $\mathcal{L}_{\mathsf{B}} \subset \mathcal{L}_{\mathsf{B'}}$. Therefore, conservativity is again preserved.

[13] To be more precise, one can formulate a formal conditions in order to ensure that relative truth definability generally entails proof-theoretic reducibility. See Fujimoto [22], p. 325.

$$\begin{tikzcd}
\mathsf{PUTB} \arrow[r, Leftrightarrow] & \mathsf{KF} & \mathbf{DT} \arrow[l, Lleftarrow] & & \mathsf{RT}_{\epsilon_0} \arrow[ll, Leftarrow] \\
& & & & \mathsf{CT} \arrow[u, Rightarrow] \\
\mathsf{PUTB}\upharpoonright \arrow[uu, Rightarrow] \arrow[r, Leftrightarrow] & \mathsf{KF}\upharpoonright \arrow[uu, Rightarrow] & \mathbf{DT}\upharpoonright \arrow[uu, Rightarrow] \arrow[l, Lleftarrow] & \mathsf{UTB} \arrow[l, Leftarrow] \arrow[r, Rightarrow] & \mathsf{CT}\upharpoonright \arrow[u, Rightarrow]
\end{tikzcd}$$

the converse holds. In what follows I will explain the proof of some of these results presented by Fujimoto in [22].

DT and RT

The first point we focus on is the relationship between DT and the theory of ramified truth RT. As a reminder the theory of ramified truth up to α ($\mathsf{RT}_{<\alpha}$) is formulated in a language $\mathcal{L}_{<\alpha}$ which is $\mathcal{L}_{\mathsf{PA}}$ expanded by all truth predicates T_β for all $\beta < \alpha$. And $\mathrm{Sent}_\beta(x)$ is a primitive recursive predicate which stands for $x \in \mathrm{Sent}_{\mathcal{L}_\beta}$. I rewrite the definition:

Definition 3.2.1. For $\alpha \leq \Gamma_0$ the theory $\mathsf{RT}_{<\alpha}$ is given by all the axioms of PA, induction axioms for $\mathcal{L}_{<\alpha}$ and, for all $\gamma < \beta < \alpha$:

(RT1) $\forall s \forall t\ [T_\beta \ulcorner s = t \urcorner \leftrightarrow s^\circ = t^\circ]$

(RT2) $\forall x\ [\mathrm{Sent}_{<\beta}(x) \rightarrow (T_\beta(\dot{\neg} x) \leftrightarrow \neg T_\beta x)]$

(RT3) $\forall x \forall y\ [\mathrm{Sent}_{<\beta}(x \dot{\vee} y) \rightarrow (T_\beta(x \dot{\vee} y) \leftrightarrow T_\beta(x) \vee T_\beta(y))]$

(RT4) $\forall x \forall y\ [\mathrm{Sent}_{<\beta}(x \dot{\rightarrow} y) \rightarrow (T_\beta(x \dot{\rightarrow} y) \leftrightarrow (T_\beta(x) \rightarrow T_\beta(y)))]$

(RT5) $\forall v \forall x\ [\mathrm{Sent}_{<\beta}(\dot{\forall} v x) \rightarrow (T_\beta(\dot{\forall} v x) \leftrightarrow \forall t T_\beta(x[t/v]))]$

(RT6) $\forall t\ [\mathrm{Sent}_{<\gamma}(t^\circ) \rightarrow (T_\beta(\dot{T}_\gamma t) \leftrightarrow T_\gamma t^\circ)]$

(RT7) $\forall t \forall \delta \prec \bar{\beta}\ [\mathrm{Sent}_{<\delta}(t^\circ) \rightarrow (T_\beta(\dot{T}_\delta t) \leftrightarrow T_\beta t^\circ)]$

We shall see that:

Theorem 3.2.1. $\mathsf{RT}_{<\epsilon_0}$ is truth-definable in DT.

This result follows from a more general lemma.
Define a formula of $\mathcal{L}_T$ $D^+(x)$:

$$D^+(x) = \mathrm{Sent}_T(x) \wedge (T(x) \vee F(x)) \wedge \neg (T(x) \wedge F(x)) \equiv \mathrm{Sent}_T(x) \wedge (F(x) \leftrightarrow \neg T(x)).$$

This is a condition of ‘strengthened determinate meaningfulness’, we shall see later its link with $D(x)$.

Lemma 3.2.1. Let Q be a theory over $\mathcal{L}_T$ which proves the following:

(i) $\forall s \forall t\ [T \ulcorner s \dot{=} t \urcorner \leftrightarrow s^\circ = t^\circ] \wedge \forall s \forall t\ [F(s = t) \leftrightarrow s^\circ \neq t^\circ]$

(ii) $[D^+(x) \wedge D^+(y)] \rightarrow [D^+(\underset{.}{\neg}x) \wedge D^+(x\underset{.}{\vee}y) \wedge D^+(x\underset{.}{\rightarrow}y)]$;

(iii) $[\mathrm{Sent}_T(\underset{.}{\forall}vx) \wedge \forall y D^+(x)[y/v]] \rightarrow D^+(\underset{.}{\forall}vx)$;

(iv) $D^+(x) \rightarrow (T(\underset{.}{\neg}x) \leftrightarrow \neg T(x))$;

(v) $D^+(x\underset{.}{\vee}y) \rightarrow (T(x\underset{.}{\vee}y) \leftrightarrow T(x) \vee T(y))$;

(vi) $D^+(x\underset{.}{\rightarrow}y) \rightarrow (T(x\underset{.}{\rightarrow}y) \leftrightarrow T(x) \rightarrow T(y))$;

(vii) $D^+(\underset{.}{\forall}vx) \rightarrow (T(\underset{.}{\forall}vx) \leftrightarrow \forall y T(x[y/v]))$;

(viii) $\forall t\ [T(\underset{.}{T}x) \leftrightarrow T(t^{\circ})] \wedge \forall t\ [F(\underset{.}{T}x) \leftrightarrow F(t^{\circ})]$.

Then, if $\mathsf{Q} \vdash \mathrm{TI}_{\mathcal{L}_T}(< \alpha)$, then $\mathsf{RT}_{<\alpha}$ is truth-definable in Q[14].

Before illustrating the proof, let us explain the conditions: (i) and (viii) are the axioms (KF1) and (KF2) of KF, i.e. respectively truth and falsity of atomic sentences of $\mathcal{L}_{\mathsf{PA}}$ and $\mathcal{L}_T$; (ii) and (iii) stand for the strong compositionality of the predicate D^+ and, moreover, (iv) – (vii) represent compositionality of T for meaningful sentences. Lastly, $\mathrm{TI}_{\mathcal{L}_T}(< \alpha)$ denotes the schema of transfinite induction up to α for formulae of $\mathcal{L}_T$, that is for all $\beta < \alpha$:

$$\forall x(\forall y < x\ \phi(y) \rightarrow \phi(x)) \rightarrow \forall x \leq \beta(\phi(x)) \text{ for all } \phi \in \mathcal{L}_T.$$

Proof. Let h be a binary primitive function such that:

$$h(x, \beta) := \begin{cases} x, & \text{if } x \in \mathrm{Sent}_\beta \text{ and } \beta < \alpha; \\ \ulcorner 0 = 1 \urcorner, & \text{otherwise.} \end{cases}$$

We write h_β for $h(x, \beta)$: for simplicity, the second argument becomes a subscript. Intuitively, we have to interpret $\mathsf{RT}_{<\alpha}$ in Q. The former has a hierarchical language with several truth predicates ($\{T_i\}_{i<\alpha}$), whereas the language of Q has just one truth predicate, T. For this reason we need a function dependent from the index β, like h_β, that somehow acts as a filter for the sentences by returning all and only the sentences belonging to the language $\mathcal{L}_\beta$. In other words, for each stage of the hierarchy we have an effective method (a test) to determine whether a sentence belongs or not to that stage, i.e. whether it contains at least one occurrence of the truth predicate that "controls" the stage. If the test result over a sentence is positive the function gives us the sentence itself.

[14] This lemma can be generalized to theories with a base theory that is not PA, in this case condition (i) must be

$$\forall t_1, \ldots, \forall t_n\ [T(\underset{.}{R}(t_1, \ldots, t_n)) \leftrightarrow R(t_1{}^{\circ}, \ldots, t_n{}^{\circ})] \wedge \forall t_1, \ldots, \forall t_n\ [F(\underset{.}{R}(t_1, \ldots, t_n)) \leftrightarrow \neg R(t_1{}^{\circ}, \ldots, t_n{}^{\circ})],$$

for each atomic R of the base theory language.

Then, by using the primitive recursion theorem, we take another primitive recursive function k such that:

$$k(a) := \begin{cases} a, & \text{if } a \in \mathrm{AtFml}_{\mathcal{L}}; \\ \ulcorner T(k \circ h_\gamma(b))\urcorner, & \text{if } a = \underset{\cdot}{T}_\gamma b \text{ for } b \in \mathrm{Term}; \\ \underset{\cdot}{\neg}\ulcorner Tk\dot{b}\urcorner, & \text{if } a = \underset{\cdot}{\neg} b; \\ \ulcorner Tk\dot{b}\urcorner \underset{\cdot}{\vee} \ulcorner Tk\dot{c}\urcorner, & \text{if } a = b \underset{\cdot}{\vee} c; \\ \ulcorner Tk\dot{b}\urcorner \underset{\cdot}{\rightarrow} \ulcorner Tk\dot{c}\urcorner, & \text{if } a = b \underset{\cdot}{\rightarrow} c; \\ \underset{\cdot}{\forall} x \ulcorner Tk(\dot{b}[x/\ulcorner x \urcorner])\urcorner, & \text{if } a = \underset{\cdot}{\forall} x b \text{ for } x \in \mathrm{Var}; \\ \ulcorner 0 = 1 \urcorner, & \text{otherwise.} \end{cases}$$

I am sloppy in the notation, for example $k(\underset{\cdot}{T}_\gamma b)$ means $T(k \circ h_\gamma(b))$ and $\dot{b}$ stands for the formula obtained by substituting the free variables in b with the numerals of fresh variables. The operation expressed by the symbol $\circ$, namely the composition between functions, is defined as usual.

The function k is defined along the inductive definition of formula. If its argument is an atomic formula of the base language, then it gives as output the formula itself. Moreover, k applies to sentences that might contains an indexed truth predicate. In these cases the hierarchical index is transferred from the truth predicate to the sentences said true, i.e. the sentence to which the predicate applies by using the test h. Sentences resulting from the application of k will be sentences of $\mathcal{L}_T$, in other words it holds the following:

$$x \in \mathrm{Sent}_\beta \rightarrow kx \in \mathrm{Sent}_T \quad \text{for each } \beta < \alpha.$$

So, the original sentences of $\{\mathcal{L}_i\}_{i<\alpha}$ are transformed by k into sentences of $\mathcal{L}_T$. The next step is to prove that for each fixed β the transformed sentences are meaningful in Q, in the strengthened sense expressed by D^+:

$$\mathsf{Q} \vdash \forall \gamma \leq \beta \ (\mathrm{Sent}_\gamma(x) \rightarrow D^+(kx)). \tag{3.2}$$

Let $\phi(\rho)$ be the formula $\mathrm{Sent}_\rho(x) \rightarrow D^+(x)$. We have to prove $\phi(\gamma)$ for each ordinal γ up to β, so a transfinite argument is required. Since $\beta < \alpha$ by hypothesis we have $\mathsf{Q} \vdash \mathrm{TI}_{\mathcal{L}_T}(\beta)$ and moreover $\phi(\gamma) \in Fml_{\mathcal{L}_T}$, then:

$$\mathsf{Q} \vdash \forall \gamma (\forall \delta < \gamma \ \phi(\delta) \rightarrow \phi(\gamma)) \rightarrow \forall \gamma \leq \beta (\phi(\gamma)).$$

Thus, in order to prove (3.2) we just need to show that ϕ is progressive on γ: we suppose the claim holds up to γ (**I.H.**) and by subinduction on the complexity of the sentences we prove the claim for γ itself. Let a be a code of a sentence of $\mathcal{L}_\gamma$.

- $a \in \mathrm{AtSent}$
  Since by definition $ka = a$, the thesis is $D^+(a)$. This follows immediately from (i):

$$D^+(a) \Leftrightarrow (Ta \vee Fa) \wedge \neg(Ta \wedge Fa) \underset{(i)}{\Leftrightarrow} (a \vee a) \wedge \neg(a \wedge a).$$

- $a \in \mathrm{AtSent}_\gamma$
  $a = \underset{\cdot}{T}_\delta(t)$ with $\delta < \gamma$. Thus $ka = \ulcorner T(k \circ h_\delta(t))\urcorner$. Then:

$$\begin{aligned} D^+(ka) \Leftrightarrow\ & [T\ulcorner T(k \circ h_\delta(t))\urcorner \vee F\ulcorner T(k \circ h_\delta(t))\urcorner] \\ & \wedge [\neg T\ulcorner T(k \circ h_\delta(t))\urcorner \vee \neg F\ulcorner T(k \circ h_\delta(t))\urcorner] \\ \underset{(viii)}{\Longleftrightarrow}\ & [T(k \circ h_\delta(t^{\circ})) \vee F(k \circ h_\delta(t^{\circ}))] \\ & \wedge [\neg T(k \circ h_\delta(t^{\circ})) \vee \neg F(k \circ h_\delta(t^{\circ}))]. \end{aligned}$$

  Now, we investigate the term $k \circ h_\delta(t^{\circ})$:

$$h_\delta(t^{\circ}) := \begin{cases} t^{\circ}, & \text{if } t^{\circ} \in \mathrm{Sent}_\delta; \qquad 1. \\ \ulcorner 0 = 1\urcorner, & \text{otherwise.} \qquad 2. \end{cases}$$

  1. $k \circ h_\delta(t^{\circ}) = k(t^{\circ})$. The thesis $D^+(k(t^{\circ}))$ follows from I.H., since $t^{\circ} \in \mathrm{Sent}_\delta$ and $\delta < \gamma$.
  2. $k \circ h_\delta(t^{\circ}) = \ulcorner 0 = 1\urcorner$. The claim trivially follows by condition (i).

For the cases in which a is a propositional compound, we prove a more general fact:

$$D^+(kx) \leftrightarrow D^{+\ulcorner}Tk\dot{x}^{\urcorner} \text{ for all } x \in \mathrm{Sent}_\gamma. \tag{3.3}$$

This is shown by:

$$\begin{aligned} D^+(kx) \Leftrightarrow\ & [Tkx \vee Fkx] \wedge [\neg Tkx \vee \neg Fkx] \\ \Leftrightarrow\ & [T\ulcorner Tk\dot{x}\urcorner \vee F\ulcorner Tk\dot{x}\urcorner] \wedge [\neg T\ulcorner Tk\dot{x}\urcorner \vee \neg F\ulcorner Tk\dot{x}\urcorner] \\ \Leftrightarrow\ & D^{+\ulcorner}Tk\dot{x}^{\urcorner}. \end{aligned}$$

The intermediate step follows from an equivalent version of (viii):

$$\forall x(T(\underset{\cdot}{T}num(x)) \leftrightarrow T(x)).$$

As a consequence, (3.3) holds. I develop just one case, being the others similar.

- $a = \underset{\cdot}{\neg} b$

$$D^+(k\underset{\cdot}{\neg}b) = D^+(\underset{\cdot}{\neg}\ulcorner Tk\dot{b}\urcorner) \underset{(ii)}{\Longleftarrow} D^{+\ulcorner}Tk\dot{b}^{\urcorner} \underset{(3.3)}{\Longleftrightarrow} D^+(kb) \text{ [S.I.H.]}$$

- $a = b\underset{\cdot}{\vee}c$ and $a = b\underset{\cdot}{\rightarrow}c$ similarly follow by (3.3) and (ii).

- $a = \underset{\cdot}{\forall}xb$
  In this case we observe a more general fact for each y in Q:

$$D^+(kb[y/x]) \Leftrightarrow D^{+\ulcorner}Tk(\dot{b}[\dot{y}/\ulcorner x\urcorner])^{\urcorner} \Leftrightarrow D^{+\ulcorner}Tk(\dot{b}[\dot{x}/\ulcorner x\urcorner])^{\urcorner}[y/x]). \tag{3.4}$$

Then, we have

$$\begin{aligned}[\text{S.I.H}]\ \forall y D^{+}k(b[y/x]) &\underset{(3.4)}{\Longleftrightarrow} \forall y D^{+}(\ulcorner Tk(\dot{b}[\dot{x}/\ulcorner x\urcorner])\urcorner[y/x])\\ &\underset{(iii)}{\Longrightarrow} D^{+}(\forall x\ulcorner Tk(\dot{b}[x/\ulcorner x\urcorner])\urcorner)\\ &= D^{+}(k\forall xb).\end{aligned}$$

We have shown (3.2).

Now we can prove the claim of the truth-definability of $\mathsf{RT}_{<\alpha}$ in Q, by defining a new formula $\theta_\beta(x)$ for each $\beta < \alpha$ which is the truth-defining formula of T_β. We set $\theta_\beta(x)$ to be $Tk(x)$ and what we are going to show is that $\theta_\beta(x)$ is the formula of $\mathcal{L}_Q$ such that

$$\mathsf{RT}_{<\alpha} \vdash \phi \Rightarrow \mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}(\phi) \quad \text{for all } \phi \in \mathcal{L}_\alpha,$$

where $\mathcal{L}_\alpha = \mathcal{L}_0 \cup \{T_\beta\}_{\beta<\alpha}$ and $\mathcal{T}_{\vec{\theta}}$ is the translation defined in 2.4.1. So, using the clauses (iv)–(viii) we show that Q proves the translated versions of the axioms of $\mathsf{RT}_{<\alpha}$.

- $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}(\text{RT1})$
  Let t and s be closed terms of $\mathcal{L}_Q$.
  Thesis: $\mathcal{T}_{\vec{\theta}}(T_\beta\ulcorner s = t\urcorner) \Leftrightarrow \mathcal{T}_{\vec{\theta}}(s^{\circ} = t^{\circ})$
  Proof:

$$\begin{aligned}\mathcal{T}_{\vec{\theta}}(T_\beta\ulcorner s = t\urcorner) &\underset{\mathcal{T}_{\vec{\theta}}}{:=} \theta_\beta\ulcorner s = t\urcorner\\ &\underset{\theta_\beta}{:=} Tk(s = t)\\ &\underset{k}{:=} T\ulcorner s = t\urcorner\\ &\underset{(i)}{\Leftrightarrow} s^{\circ} = t^{\circ}\\ &\underset{\mathcal{T}_{\vec{\theta}}}{=:} \mathcal{T}_{\vec{\theta}}(s^{\circ} = t^{\circ}).\end{aligned}$$

- $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}(\text{RT2})$
  Assume $\beta < \alpha$ and $x \in \text{Sent}_{<\beta}$
  Thesis: $\mathcal{T}_{\vec{\theta}}(T_\beta(\underset{\cdot}{\neg}x)) \Leftrightarrow \mathcal{T}_{\vec{\theta}}(\neg T_\beta x)$
  Proof:

$$\begin{aligned}\mathcal{T}_{\vec{\theta}}(T_\beta(\underset{\cdot}{\neg}x)) &\underset{\mathcal{T}_{\vec{\theta}}}{:=} \theta_\beta(\underset{\cdot}{\neg}x)\\ &\underset{\theta_\beta}{:=} Tk(\neg x)\\ &\underset{k}{:=} T(\underset{\cdot}{\neg}\ulcorner Tk\dot{x}\urcorner)\end{aligned}$$

Now we observe that for (3.2) D^+kx holds since $x \in \mathrm{Sent}_{<\beta}$. From this by (3.3) and (ii) we also have $D^+(\underset{\cdot}{\neg}\ulcorner Tk\dot{x}\urcorner)$. Thus, we can apply (iv):

$$\begin{aligned}
T(\underset{\cdot}{\neg}\ulcorner Tk\dot{x}\urcorner) &\underset{(iv)}{\Longleftrightarrow} \neg T\ulcorner Tk\dot{x}\urcorner \\
&\underset{(viii)}{\Longleftrightarrow} \neg Tkx \\
&\underset{\theta_\beta}{=:} \neg\theta_\beta(x) \\
&\underset{\mathcal{T}_{\vec{\theta}}}{=:} \neg\mathcal{T}_{\vec{\theta}}(T_\beta x) \\
&\underset{\mathcal{T}_{\vec{\theta}}}{=:} \mathcal{T}_{\vec{\theta}}(\neg T_\beta x).
\end{aligned}$$

- $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}$(RT3)
  Let x and y be sentences of $\mathcal{L}_{<\beta}$
  Thesis: $\mathcal{T}_{\vec{\theta}}(T_\beta(x \vee y)) \Leftrightarrow \mathcal{T}_{\vec{\theta}}(T_\beta x \vee T_\beta y)$
  Proof:

$$\begin{aligned}
\mathcal{T}_{\vec{\theta}}(T_\beta(x \vee y)) &\underset{\mathcal{T}_{\vec{\theta}}}{:=} \theta_\beta(x \vee y) \\
&\underset{\theta_\beta}{:=} Tk(x \vee y) \\
&\underset{k}{:=} T\ulcorner Tk\dot{x}\urcorner \underset{\cdot}{\vee} \ulcorner Tk\dot{y}\urcorner \\
&\underset{(v)}{\Leftrightarrow} T\ulcorner Tk\dot{x}\urcorner \underset{\cdot}{\vee} T\ulcorner Tk\dot{y}\urcorner \\
&\underset{(viii)}{\Longleftrightarrow} Tkx \underset{\cdot}{\vee} Tky \\
&\underset{\theta_\beta}{=:} \theta_\beta(x) \underset{\cdot}{\vee} \theta_\beta(y) \\
&\underset{\mathcal{T}_{\vec{\theta}}}{=:} \mathcal{T}_{\vec{\theta}}(T_\beta x) \underset{\cdot}{\vee} \mathcal{T}_{\vec{\theta}}(T_\beta y) \\
&\underset{\mathcal{T}_{\vec{\theta}}}{=:} \mathcal{T}_{\vec{\theta}}(T_\beta x \vee T_\beta y).
\end{aligned}$$

Again, the step in which (v) is used is justified by the following argument:

$$\begin{gathered}
x, y \in \mathrm{Sent}_{<\beta} \underset{(3.2)}{\Longrightarrow} D^+(kx) \text{ and } D^+(ky) \underset{(3.3)}{\Longrightarrow} \\
D^{+\ulcorner}Tk\dot{x}\urcorner \text{ and } D^{+\ulcorner}Tk\dot{y}\urcorner \underset{(ii)}{\Longrightarrow} D^{+\ulcorner}Tk\dot{x}\urcorner \underset{\cdot}{\vee} \ulcorner Tk\dot{y}\urcorner.
\end{gathered}$$

- $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}$(RT4) Analogous.
- $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}$(RT5) Analogous.

- $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}$(RT6)
  Let $\gamma < \beta < \alpha$ and t a term such that $t^{\circ} \in \text{Sent}_{<\gamma}$.
  Thesis: $\mathcal{T}_{\vec{\theta}}(T_\beta(\underset{\cdot}{T}_\gamma t)) \Leftrightarrow \mathcal{T}_{\vec{\theta}}(T_\gamma t^{\circ})$

$$\begin{aligned}
\mathcal{T}_{\vec{\theta}}(T_\beta(\underset{\cdot}{T}_\gamma t)) &:= \theta_\beta(T_\beta(\underset{\cdot}{T}_\gamma t)) \\
&:= Tk(T_\beta(\underset{\cdot}{T}_\gamma t)) \\
&:= T\ulcorner T(k \circ h_\beta(\underset{\cdot}{T}_\gamma t))\urcorner \\
&\underset{h_\beta}{:=} T\ulcorner T(k\underset{\cdot}{T}_\gamma t)\urcorner \\
&:= T\ulcorner T\ulcorner T(k \circ h_\gamma t)\urcorner\urcorner \\
&\underset{(viii)}{\Longleftrightarrow} T(k \circ h_\gamma t^{\circ}) \\
&\underset{h_\gamma}{:=} T(kt^{\circ}) \\
&=: \theta_\gamma(t^{\circ}) \\
&=: \mathcal{T}_{\vec{\theta}}(T_\gamma t^{\circ}).
\end{aligned}$$

- $\mathsf{Q} \vdash \mathcal{T}_{\vec{\theta}}$(RT7) Analogous to (RT6): one can take an arbitrary δ in place of γ and in the last steps put $T(kt^{\circ}) = \theta_\beta(t^{\circ})$ obtaining, thus, $\mathcal{T}_{\vec{\theta}}(T_\beta t^{\circ})$.

□

Lemma 3.2.2. DT meets conditions (i)–(viii) and moreover $\mathsf{DT} \vdash \text{TI}_{\mathcal{L}_T}(<\epsilon_0)$.

Proof. As a preliminary remark I investigate the relation between

$$D(x) = T(x) \vee F(x)$$

and

$$D^+ = \text{Sent}_T(x) \wedge (T(x) \vee F(x)) \wedge \neg(T(x) \wedge F(x)) \equiv \text{Sent}_T(x) \wedge (F(x) \leftrightarrow \neg T(x)).$$

As a reminder, I write the axiom (DT9) using the abbreviation Fx for $T\neg x$:

$$\forall x\ [\text{Sent}_T(x) \wedge D(x) \to (F(x) \leftrightarrow \neg T(x))].$$

This means that in the system DT the third clause of D^+ is entailed from the first and the second, so it can be dropped. That it to say:

$$\mathsf{DT} \vdash \forall x(\text{Sent}_T(x) \wedge D(x) \to D^+(x)).$$

And, trivially, the other direction holds as well. After this clarification is immediate to observe:

(i) (DT7);

(ii) (DT3), (DT4), (DT5);

(iii) (DT6);

(iv) (DT9);

(v) (DT10);

(vi) (DT11);

(vii) (DT12);

(viii) (DT8).

Maybe the less immediate is (DT5), nevertheless axiom (DT5) entails condition (ii) since

$$D(x) \wedge D(y) \Rightarrow D(x) \wedge (T(x) \rightarrow D(y)) \Leftrightarrow D(x \mathbin{\underset{\cdot}{\rightarrow}} y).$$

Moreover, transfinite induction up to α for each $\alpha < \epsilon_0$ can be established in DT for all formulae of $\mathcal{L}_T$[15]. □

From lemmata 3.2.1 and 3.2.2 immediately theorem 3.2.1 follows. Note that the same holds for KF: $\mathsf{RT}_{<\epsilon_0}$ is truth-definable in KF. Furthermore, neither $\mathsf{DT}\!\upharpoonright$ nor $\mathsf{KF}\!\upharpoonright$ are enough to define the truth of $\mathsf{RT}_{<\epsilon_0}$. This follows from a result of non-conservatvity and from the fact that the notion of truth-definability implies conservativity for truth-free theorems: any theory non conservative for $\mathcal{L}_{\mathsf{PA}}$ sentences over a theory Q is non truth-definable in Q. This yields the following negative results follow:

$$\mathsf{DT}\upharpoonright, \mathsf{KF}\upharpoonright \not\preceq \mathsf{RT}_{<\epsilon_0}.$$

Let us now investigate whether the reverse holds, namely is the truth predicate of DT definable in a ramified theory of truth? The answer is negative, that is the reverse fails as the following theorem states:

Theorem 3.2.2. If $\alpha \leq \epsilon_0$, then

$$\mathsf{DT}(\mathsf{KF}) \not\preceq \mathsf{RT}_{<\alpha}.$$

Proof. Let $\alpha < \epsilon_0$. Suppose $\mathsf{DT} \preceq \mathsf{RT}_{<\alpha}$. From theorem 3.2.1 we have $\mathsf{RT}_{<\epsilon_0} \preceq \mathsf{DT}$. By the transitivity of the relation of truth-definability from $\mathsf{RT}_{<\epsilon_0} \preceq \mathsf{DT}$ and $\mathsf{DT} \preceq \mathsf{RT}_{<\alpha}$ we get $\mathsf{RT}_{<\epsilon_0} \preceq \mathsf{RT}_{<\alpha}$ for all $\alpha < \epsilon_0$. This is impossible as for $\beta < \gamma$, $\mathsf{RT}_{<\gamma}$ is not conservative over $\mathsf{RT}_{<\beta}$ (indeed the former proves the consistency of the latter) and, so, since $\alpha < \epsilon_0$, $\mathsf{RT}_{<\epsilon_0} \not\preceq \mathsf{RT}_{<\alpha}$.

[15] Cfr. Feferman [18], p. 212.

Let $\alpha = \epsilon_0$. Again, for the sake of contradiction, suppose $\mathsf{DT} \preceq \mathsf{RT}_{<\epsilon_0}$. By definition of truth-definability there is a formula $\theta \in \mathcal{L}_{\epsilon_0}$ that defines the truth predicate of DT in $\mathsf{RT}_{<\epsilon_0}$, i.e.

$$\mathsf{RT}_{<\epsilon_0} \vdash \mathcal{T}_\theta(\mathrm{DT1}) \wedge \cdots \wedge \mathcal{T}_\theta(\mathrm{DT12}).$$

By the theorem of finiteness of premises there exists a finite set of formulae $\Gamma \subset \mathsf{RT}_{<\epsilon_0}$ such that:

$$\Gamma \vdash \mathcal{T}_\theta(\mathrm{DT1}) \wedge \cdots \wedge \mathcal{T}_\theta(\mathrm{DT12}).$$

Take the maximum $\beta < \epsilon_0$ such that T_β occurs in $\Gamma \cup \{\theta\}$. Formulae containing this predicate belong to the language of the next hierarchical level: $\Gamma \subset \mathsf{RT}_{<\beta+1}$. Hence:

$$\mathsf{RT}_{<\beta+1} \vdash \mathcal{T}_\theta(\mathrm{DT1}) \wedge \cdots \wedge \mathcal{T}_\theta(\mathrm{DT12}),$$

that is $\mathsf{DT} \preceq \mathsf{RT}_{<\beta+1}$. But since $\beta + 1 < \epsilon_0$ this possibility has been already ruled out. □

In this case, relative truth-definability is one-way.

Relative truth-definability assigns to theories of ramified truth a precise and well-delimitate role in intertheoretical relations:

- On the one hand, they serve as touchstone for type-free theories such as KF and DT. Systems of ramified truth can be embedded into type-free theories: those theories can 'simulate' all the truth predicates of ramified theories up to a certain level. The higher is the level of the Tarski's hierarchy reached in a theory, the stronger is the theory itself, e.g $\mathsf{RT}_{<\epsilon_0} \preceq \mathsf{DT}, \mathsf{KF}$, whereas FS defines all the truth predicate of RT up to ω[16].

- On the other hand, ramified theories of truth cannot define the truth predicate of a type-free theory even when much levels are included. Furthermore, even if the type-free theory is taken in its weaker version without induction on $\mathcal{L}_T$-formulae thy could not express its truth predicate:

  $$\mathsf{DT} \restriction, \mathsf{KF} \restriction \not\preceq \mathsf{RT}_{<\epsilon_0}.$$

  It is worth noting that this does not exclude other kinds of reduction from type-free theories to ramified theories.

Hence, it seems that relative truth-definability is able to capture a difference between ramified and type-free truth theories. For example the relative interpretability of $\mathsf{RT}_{<\omega}$ in FS also holds in the converse direction, while none of theories introduced above can define the truth if FS since no ω-consistent theory can define the truth of an ω-inconsistent theory.

[16] For a proof see Halbach [28], theorem 14.26 and corollary 14.27.

DT and KF

After having seen DT and KF united in their relations with ramified theories of truth, let us now delve how they are related to each other. Again I rewrite the definition:

Definition 3.2.2. The system KF is given by all axioms of PAT and the following axioms:

(KF1) $\forall s \forall t\ [T\ulcorner s \underset{\cdot}{=} t\urcorner \leftrightarrow s^{\circ} = t^{\circ}] \land \forall s \forall t\ [F\ulcorner s \underset{\cdot}{=} t\urcorner \leftrightarrow s^{\circ} \neq t^{\circ}]$

(KF2) $\forall t\ [T(\underset{\cdot}{T}t) \leftrightarrow Tt^{\circ}] \land \forall t\ [F(\underset{\cdot}{T}t) \leftrightarrow (T\neg t^{\circ})]$

(KF3) $\forall x\ [\mathrm{Sent}_T(x) \rightarrow (T(\underset{\cdot}{\neg}\underset{\cdot}{\neg}x) \leftrightarrow Tx)]$

(KF4) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\vee} y) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow T(x) \vee T(y))]$

(KF5) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\vee} y) \rightarrow (F(x \underset{\cdot}{\vee} y) \leftrightarrow F(x) \land F(y))]$

(KF6) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\rightarrow} y) \rightarrow (T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow T(\underset{\cdot}{\neg}x \underset{\cdot}{\vee} y))]$

(KF7) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\rightarrow} y) \rightarrow (F(x \underset{\cdot}{\rightarrow} y) \leftrightarrow F(\underset{\cdot}{\neg}x \underset{\cdot}{\vee} y))]$

(KF8) $\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\forall} vx) \rightarrow (T(\underset{\cdot}{\forall} vx) \leftrightarrow \forall t T(x[t/v]))]$

(KF9) $\forall v \forall x\ [\mathrm{Sent}_T(\underset{\cdot}{\forall} vx) \rightarrow (F(\underset{\cdot}{\forall} vx) \leftrightarrow \exists t F(x[t/v]))]$

With respect to the definition in chapter 1, axioms (KF6) and (KF7) are added. They are redundant since → is definable by ¬ and ∨ in Strong Kleene Logic, but this formulation will be useful in view of a comparison with DT in which → is taken as a primitive connective. Apparently, DT and KF are very different both for their axiomatizations and their background motivations; but we shall see that at a closer look we find they are more related than they appear. Leaving aside the philosophical ideas behind the two systems — I shall come back to them later — let us focus on formal obstacles to a possible reduction. It seems that there are essentially two of them:

(i) The inner logic: SKL for KF and a variant of WKL for DT.

(ii) The relation between 'to be true' and 'to be false': in DT for sentences satisfying D saying that x is not true is the same of saying that x is false ((DT9)); whereas in KF 'to be not true' and 'to be false' are deliberately kept separate, that is why axioms are split in two.

In order to bridge the first gap, Fujimoto builds a variant of KF, i.e. an iterative compositional theory, whose axioms for compositionality are given according to what he calls Feferman Logic, that is the variant of weak Kleene Logic we have seen before. In this logic the conditional → cannot be defined in terms of $\neg$ and $\vee$. The axioms that should be changed are (KF4), (KF6), (KF7) and (KF9), that is respectively the axioms which govern the truth of a disjunction, the truth and the falsity of a conditional and lastly the falsity of a sentence of the form $\forall xb$. The universal quantifier is interpreted as infinitary conjunction (a weak Kleene one), thus a sentence like $\forall xb$ is false not only if $b(x)$ is false for some x, but even if $b(x)$ is either true or false for all x.

Definition 3.2.3. The system $\mathsf{FKF}\restriction$ is given by all axioms of PA and (KF1) – (KF3), (KF5), (KF8) and:

(FKF4) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\vee} y) \to (T(x \underset{\cdot}{\vee} y) \leftrightarrow ((T(x) \wedge T(y)) \vee (F(x) \wedge T(y)) \vee$
$\vee\ (T(x) \wedge F(y))))]$

(FKF6) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\to} y) \to (T(x \underset{\cdot}{\to} y) \leftrightarrow ((T(x) \wedge T(y)) \vee F(x)))]$

(FKF7) $\forall x \forall y\ [\mathrm{Sent}_T(x \underset{\cdot}{\to} y) \to (F(x \underset{\cdot}{\to} y) \leftrightarrow (T(x) \wedge F(y)))]$

(FKF9) $\forall v \forall x\ [\mathrm{Sent}_T(\forall vx) \to (F(\underset{\cdot}{\forall} vx) \leftrightarrow (\forall t(T(x[t/v]) \vee F(x[t/v])) \wedge$
$\wedge\ \exists t F(x[t/v]))))]$

As usual, FKF is $\mathsf{FKF}\restriction$ plus full induction for $\mathcal{L}_T$.

This theory acts somehow as a bridge between the others: we shall show that FKF and DT are equivalent, then that the truth of FKF is definable in KF. FKF can play this role by virtue of two key features: it is an iterative compositional theory (just like KF) and, moreover, shares with DT the 'inner logic'.

Theorem 3.2.3. DT ($\mathsf{DT}\restriction$) and FKF + Cons ($\mathsf{FKF}\restriction$ + Cons) are identical theories.

As a reminder, the axiom Cons is

$$\forall x\ [\mathrm{Sent}_T(x) \to \neg(T(x) \wedge F(x))].$$

It says that no sentence is true and false, in other words it rejects truth-value gluts. If we look at the alternative formulation of this axiom introduced by Halbach[17], i.e

$$\forall x\ [\mathrm{Sent}_T(x) \to (T(\underset{\cdot}{\neg} x) \to \neg T(x))]$$

it becomes immediately clear its similarity with (DT9). So, the presence of Cons partially obviates the problem (ii), for this reason Cons is a key player in the construction of a theory akin to KF but identical to DT.

[17] [28], p. 155.

Proof. The proof is given by showing that all the axioms of one are provable in the other. We shall see just some examples. Let us reason informally in DT. In order to show $\mathsf{DT} \vdash \mathsf{Cons}$, the strategy is to assume the negation of Cons together with DT e get contradiction. Let x and y be (numeral of codes of) sentences of $\mathcal{L}_T$. If $T(x) \wedge F(x)$ then $D(x)$. For (DT9), $F(x) \leftrightarrow \neg T(x)$, that is $(T(x) \vee F(x)) \vee \neg(T(x) \wedge F(x))$, which contradicts our assumption. Thus $\mathsf{DT} \vdash \mathsf{Cons}$. As further example we prove that $\mathsf{FKF}+\mathsf{Cons} \vdash$ (DT10). Given any x, $y \in \mathrm{Sent}_T$ if $T(x \underset{\cdot}{\vee} y)$ then $(T(x) \wedge T(y)) \vee (F(x) \wedge T(y)) \vee (T(x) \wedge F(y))$ for (FKF4). This entail $T(x) \vee T(y)$. Conversely, assume $T(x) \vee T(y)$ and $D(x \underset{\cdot}{\vee} y)$. From the second, by (FKF4) and (KF5), it follows $T(x \underset{\cdot}{\vee} y) \vee F(x \underset{\cdot}{\vee} y) \Leftrightarrow (T(x) \wedge T(y)) \vee (F(x) \wedge T(y)) \vee (T(x) \wedge F(y)) \vee (F(x) \vee F(y))$. But the last clause has to be dropped because as hypothesis we have $T(x) \vee T(y)$, thus we get $T(x \underset{\cdot}{\vee} y)$.

□

In order to prove that DT is truth-definable in KF, it is enough to show that the relation holds between the same theories without T-induction. That is because truth-definability has an important and technically useful property:

Remark 3.2.1. Suppose Q is truth-definable in S. Then, Q plus full induction for $\mathcal{L}_{\mathsf{Q}}$ is truth-definable in S plus full induction for $\mathcal{L}_{\mathsf{Q}}$.

As far as Q and S are truth theories over PA we have:

$$\text{if } \mathsf{Q} \upharpoonright \preceq \mathsf{S} \upharpoonright \text{ then } \mathsf{Q} \preceq \mathsf{S}.$$

So, by exploiting this property we can restrict to show that $\mathsf{DT}\upharpoonright$ is truth-definable in $\mathsf{KF}\upharpoonright$.

Theorem 3.2.4. $\mathsf{DT}\upharpoonright \preceq \mathsf{KF}\upharpoonright$.

Proof. We provide a formula θ which defines the truth of $\mathsf{FKF}\upharpoonright + \mathsf{Cons}$ in $\mathsf{KF}\upharpoonright$. First of all, we define a translation I in PA by using the recursion theorem. For the seek of simplicity, let us assume that I acts on the codes of formulae. We write $\underset{\cdot}{I}$ for the representation of I.

$$I(\phi) := \begin{cases} \phi, & \text{if } \phi \in \mathrm{AtFml}_{\mathcal{L}}; \\ T\underset{\cdot}{I}t, & \text{if } \phi = T(t); \\ T\underset{\cdot}{I}\underset{\cdot}{\neg}t, & \text{if } \phi = \neg T(t); \\ I(\psi), & \text{if } \phi = \neg\neg\psi; \\ (I\psi_0 \wedge I\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\neg\psi_1) & \\ \quad \vee(I\neg\psi_0 \wedge I\psi_1 \wedge \neg I\psi_0 \wedge \neg I\neg\psi_1) & \\ \quad \vee(I\psi_0 \wedge I\neg\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\psi_1), & \text{if } \phi = \psi_0 \vee \psi_1; \\ I\neg\psi_0 \wedge I\neg\psi_1 \wedge \neg I\psi_0 \wedge \neg I\psi_1, & \text{if } \phi = \neg(\psi_0 \vee \psi_1); \\ (I\psi_0 \wedge I\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\neg\psi_1) & \\ \quad \vee(I\neg\psi_0 \wedge \neg I\psi_0), & \text{if } \phi = \psi_0 \to \psi_1; \\ I\psi_0 \wedge I\neg\psi_1 \wedge \neg I\neg\psi_0 \neg I\psi_1, & \text{if } \phi = \neg(\psi_0 \to \psi_1); \\ \forall x(I\psi \wedge \neg I\neg\psi), & \text{if } \phi = \forall x\psi; \\ \forall x((I\psi \wedge \neg I\neg\psi) \wedge (I\neg\psi \wedge \neg I\psi)) & \\ \quad \wedge\exists x I\neg\psi, & \text{if } \phi = \forall x\psi; \\ 0, & \text{otherwise.} \end{cases}$$

Then define:

$$\theta(x) = TIx \wedge \neg TI\underset{\cdot}{\neg}x \wedge FI\underset{\cdot}{\neg}x \wedge \neg FIx.$$

$\theta(x)$ is the truth-defining formula we are looking for. As we did for the lemma 3.2.1, we have to show that $\mathsf{KF}{\upharpoonright}$ proves the translated version of the axioms of $\mathsf{FKF}{\upharpoonright} + \mathsf{Cons}$. We shall show some typical case:

- $\mathsf{KF}{\upharpoonright} \vdash \mathcal{T}_\theta(\text{KF1})$
  Let assume $\mathcal{T}_\theta(T\ulcorner s\underset{\cdot}{=}t\urcorner)$. By definition of $\mathcal{T}_\theta$ we have $\theta(s = t)$. Note that $I\phi = \phi$ and $\mathcal{T}_\theta(\phi) = \phi$ when ϕ is an atomic sentence of the base language. Thus:

  $$\begin{aligned} \mathcal{T}_\theta(T\ulcorner s\underset{\cdot}{=}t\urcorner) &:= T\ulcorner s\underset{\cdot}{=}t\urcorner \wedge \neg T\underset{\cdot}{\neg}\ulcorner s\underset{\cdot}{=}t\urcorner \wedge F\underset{\cdot}{\neg}\ulcorner s\underset{\cdot}{=}t\urcorner \wedge \neg F\underset{\cdot}{\neg}\ulcorner s\underset{\cdot}{=}t\urcorner \\ &\underset{\text{(KF1)}}{\Longleftrightarrow} s^{\circ} = t^{\circ} \wedge s^{\circ} = t^{\circ} \wedge s^{\circ} = t^{\circ} \wedge s^{\circ} = t^{\circ} \\ &\Leftrightarrow s^{\circ} = t^{\circ} \\ &=: \mathcal{T}_\theta(s^{\circ} = t^{\circ}). \end{aligned}$$

  We reason, quite informally, in $\mathsf{KF}{\upharpoonright}$ so we can use one of its axiom, also (KF1), the translated version of which we are proving. The clause for F can be shown similarly.

- $\mathsf{KF}{\upharpoonright} \vdash \mathcal{T}_\theta(\text{KF2})$

$$\begin{aligned}
\mathcal{T}_\theta(T(\underset{\cdot}{T}t)) &:= \theta(\underset{\cdot}{T}t)\\
&:= TI\underset{\cdot}{T}t \wedge \neg TI\underset{\cdot}{\neg}\underset{\cdot}{T}t \wedge FI\underset{\cdot}{\neg}\underset{\cdot}{T}t \wedge \neg FI\underset{\cdot}{T}t\\
&:= T(\underset{\cdot}{T}It) \wedge \neg T(\underset{\cdot}{T}I\underset{\cdot}{\neg}t) \wedge F(\underset{\cdot}{T}I\underset{\cdot}{\neg}t) \wedge \neg F(\underset{\cdot}{T}It)\\
&\underset{\text{(KF2)}}{\Longleftrightarrow} TIt^{\circ} \wedge \neg TI\underset{\cdot}{\neg}t^{\circ} \wedge FI\underset{\cdot}{\neg}t^{\circ} \wedge \neg FIt^{\circ}\\
&=: \theta(t^{\circ})\\
&=: \mathcal{T}_\theta(Tt^{\circ}).
\end{aligned}$$

In the same way it can be shown in $\mathsf{KF}\!\restriction$ that $\mathcal{T}_\theta(F(\underset{\cdot}{T}t)) \leftrightarrow \mathcal{T}_\theta(Ft^{\circ})$.

- $\mathsf{KF}\!\restriction\, \vdash \mathcal{T}_\theta$(FKF4) Let $x\underset{\cdot}{\vee}y$ be (the code of) a sentence of $\mathcal{L}_T$.

$$\begin{aligned}
&\mathcal{T}_\theta(T(x\underset{\cdot}{\vee}y))\\
&:= \theta(x\underset{\cdot}{\vee}y)\\
&:= TI(x\underset{\cdot}{\vee}y) \wedge \neg TI\underset{\cdot}{\neg}(x\underset{\cdot}{\vee}y) \wedge FI\underset{\cdot}{\neg}(x\underset{\cdot}{\vee}y) \wedge \neg FI(x\underset{\cdot}{\vee}y)\\
&:= T[(Ix \wedge Iy \wedge \neg I\underset{\cdot}{\neg}x \wedge \neg I\underset{\cdot}{\neg}y) \vee (I\underset{\cdot}{\neg}x \wedge Iy \wedge \neg Ix \wedge \neg I\underset{\cdot}{\neg}y)\\
&\qquad\qquad\qquad\qquad \vee (Ix \wedge I\underset{\cdot}{\neg}y \wedge \neg I\underset{\cdot}{\neg}x \wedge \neg Iy)]\\
&\quad \wedge \neg T[I\underset{\cdot}{\neg}x \wedge I\underset{\cdot}{\neg}y \wedge \neg Ix \wedge \neg Iy]\\
&\quad \wedge F[I\underset{\cdot}{\neg}x \wedge I\underset{\cdot}{\neg}y \wedge \neg Ix \wedge \neg Iy]\\
&\quad \wedge \neg F[(Ix \wedge Iy \wedge \neg I\underset{\cdot}{\neg}x \wedge \neg I\underset{\cdot}{\neg}y) \vee (I\underset{\cdot}{\neg}x \wedge Iy \wedge \neg Ix \wedge \neg I\underset{\cdot}{\neg}y)\\
&\qquad\qquad\qquad\qquad \vee (Ix \wedge I\underset{\cdot}{\neg}y \wedge \neg I\underset{\cdot}{\neg}x \wedge \neg Iy)]\\
&\underset{*}{\Leftrightarrow} [(TIx \wedge TIy \wedge FI\underset{\cdot}{\neg}x \wedge FI\underset{\cdot}{\neg}y) \vee (TI\underset{\cdot}{\neg}x \wedge TIy \wedge FIx \wedge FI\underset{\cdot}{\neg}y)\\
&\qquad\qquad\qquad\qquad \vee (TIx \wedge TI\underset{\cdot}{\neg}y \wedge FI\underset{\cdot}{\neg}x \wedge FIy)]\\
&\quad \wedge [\neg TI\underset{\cdot}{\neg}x \vee \neg TI\underset{\cdot}{\neg}y \vee \neg FIx \wedge \neg FIy]\\
&\quad \wedge [FI\underset{\cdot}{\neg}x \vee FI\underset{\cdot}{\neg}y \vee TIx \vee TIy]\\
&\quad \wedge [(\neg FIx \wedge \neg FIy \wedge \neg TI\underset{\cdot}{\neg}x \wedge \neg TI\underset{\cdot}{\neg}y)\\
&\qquad\qquad\qquad\qquad \vee (\neg FI\underset{\cdot}{\neg}x \wedge \neg FIy \wedge \neg TIx \wedge \neg TI\underset{\cdot}{\neg}y)\\
&\qquad\qquad\qquad\qquad \vee (\neg FIx \wedge \neg FI\underset{\cdot}{\neg}y \wedge \neg TI\underset{\cdot}{\neg}x \wedge \neg TIy)].
\end{aligned}$$

The step (*) is justified by (KF3), (KF4), (KF5). In the second and third clauses we must turn $\wedge$ in $\vee$ because just one of those conditions is enough to falsify (or to make not true) a conjunction. Now, we observe that the the second and the third clauses are entailed from the fourth and first, so they can be dropped. Then, modulo logical equivalence we get:

$$\begin{aligned}
&(TIx \wedge TIy \wedge FI\underset{\cdot}{\neg}x \wedge FI\underset{\cdot}{\neg}y \wedge \neg FIx \wedge \neg FIy \wedge \neg TI\underset{\cdot}{\neg}x \wedge \neg TI\underset{\cdot}{\neg}y)\\
&\vee (TI\underset{\cdot}{\neg}x \wedge TIy \wedge FIx \wedge FI\underset{\cdot}{\neg}y \wedge \neg FI\underset{\cdot}{\neg}x \wedge \neg FIy \wedge \neg TIx \wedge \neg TI\underset{\cdot}{\neg}y)\\
&\vee (TIx \wedge TI\underset{\cdot}{\neg}y \wedge FI\underset{\cdot}{\neg}x \wedge FIy \wedge \neg FIx \wedge \neg FI\underset{\cdot}{\neg}y \wedge \neg TI\underset{\cdot}{\neg}x \wedge \neg TIy)\\
&=: (\theta(x) \wedge \theta(y)) \vee (\theta(\neg x) \wedge \theta(y)) \vee (\theta(x) \wedge \theta(\neg y))\\
&=: \mathcal{T}_\theta((T(x) \wedge T(y)) \vee (F(x) \wedge T(y)) \vee (T(x) \wedge F(y))).
\end{aligned}$$

- $\mathsf{KF}\!\restriction\ \vdash \mathcal{T}_\theta(\mathsf{Cons})$ Let us prove that $\mathcal{T}_\theta(Tx \wedge T\underset{\cdot}{\neg}x)$ is not provable in $\mathsf{KF}\!\restriction$.

$$\begin{aligned}
\mathcal{T}_\theta(Tx \wedge T\underset{\cdot}{\neg}x) &:= (\theta(x) \wedge \theta(\neg x))\\
&:= (TIx \wedge \neg TI\underset{\cdot}{\neg}x \wedge FI\underset{\cdot}{\neg}x \wedge \neg FIx\\
&\qquad \wedge TI\underset{\cdot}{\neg}x \wedge \neg TIx \wedge FIx \wedge \neg FI\underset{\cdot}{\neg}x).
\end{aligned}$$

By shuffling the members we get:

$$(TIx \wedge \neg TIx) \wedge (\neg TI\underset{\cdot}{\neg}x \wedge TI\underset{\cdot}{\neg}x) \wedge (FI\underset{\cdot}{\neg}x \wedge \neg FI\underset{\cdot}{\neg}x) \wedge (\neg FIx \wedge FIx),$$

which is of course not provable in $\mathsf{KF}\!\restriction$, whose outer logic is classical.

□

From theorem 3.2.4 and remark 3.2.1, we get:

Theorem 3.2.5. DT is truth-definable in KF.

Although the result is unique, it is obtained in two distinct steps which ought to be analysed individually.

- Equivalence between DT and FKF + Cons. There are no problems about compositionality since they share the same inner logic. Moreover, for sentences satisfying D the underlying logic of DT is consistent and complete, as the axiom (DT9) states. This easily allows the chance of proving all the axioms of the former in the latter and *vice versa.*
- Truth definability of FKF + Cons in KF. This is less unexpected since KF and FKF + Cons differ to each other just for the evaluation schema. In order to translate the latter in the former it is enough to build in KF a function (I) which simulates compositional clauses of FKF. Secondly, Cons is balanced by choosing a truth-defining formula θ that forces a true sentence to not be false and a non true sentence to be false.

The direction of the reduction (from DT to KF) is somehow natural: it shows that theories with a weaker evaluation schema are reducible to those with a stronger schema. What about the converse direction? Whether DT defines the truth of KF is an open problem. The solution might be achieved by resolving one of the following:

Open problem. Is KF ($\mathsf{KF}{\restriction}$) truth-definable in FKF ($\mathsf{FKF}{\restriction}$) + Cons?

Open problem. Is PUTB ($\mathsf{PUTB}{\restriction}$) truth-definable in DT ($\mathsf{DT}{\restriction}$)?

In the second case, the relative truth definability of KF in DT would follow by transitivity from the fact that

$\mathsf{KF} \preceq \mathsf{PUTB}$[18].

We may wonder whether such reduction is desirable since the theories at the stake are ultimately different in some respects, but a possible reduction might show that despite differences they do not embody incompatible conceptions of truth.

3.2.2 Strength

At the end of [18], Feferman raised the question of the proof-theoretic strength of DT, that he guessed to be the same as that of $\mathsf{RA}_{<\epsilon_0}$, namely Ramified Analysis up to the ordinal ϵ_0.

Feferman's First Conjecture. $\mathsf{DT} \equiv \mathsf{RA}_{<\epsilon_0}$.

Although Feferman suggested a model-theoretic strategy for the proof, Kentaro Fujimoto proved this conjecture by another, indirect, strategy: as we have seen in the previous section, he has found an interpretation of $\mathsf{RT}_{<\epsilon_0}$ in DT and of DT in KF, thus confirming the guess.

Frequently, in order to analyse the strength of a theory of truth it is related to other systems. As the following diagram shows, this is the case:

$$\begin{array}{ccccc} \mathsf{RT}_{<\epsilon_0} & \preceq & \mathsf{DT} & \preceq & \mathsf{KF} \\ \equiv & & & & \equiv \\ \mathsf{RA}_{<\epsilon_0} & & & & \mathsf{RA}_{<\epsilon_0} \\ \text{lower bound} & & & & \text{upper bound} \end{array}$$

We have already proved the following:

Theorem 3.2.6. $\mathsf{RT}_{<\epsilon_0}$ is truth-definable in DT.

Moreover, we know that:

Theorem 3.2.7. $\mathsf{RT}_{<\alpha} \equiv \mathsf{RA}_{<\alpha}$ for each ordinal α.

[18] For a proof of this result see Halbach [27], p. 792.

These theorems entail that $\mathsf{RA}_{<\epsilon_0}$ is the **lower bound** for the strength of DT. This results confirms the relevance of ramified theories of truth as touchstone for other theories: they are used to measure the strength of type-free theories. As we have seen, this is done by investigating how many stages of ramified truth can be defined in the type-free systems.

For the upper bound, there is an interpretation of DT in KF:

Theorem 3.2.8. DT is truth-definable in KF.

Moreover, it is well-known that:

Theorem 3.2.9. $\mathsf{KF} \equiv \mathsf{RA}_{<\epsilon_0}$.

These theorems entail that $\mathsf{RA}_{<\epsilon_0}$ is the **upper bound** for the strength of DT.

As a conclusion the theory DT is proof-theoretically equivalent to $\mathsf{RA}_{<\epsilon_0}$ so that Feferman's first conjecture about the proof-theoretical power of DT is verified.

$\mathsf{RA}_{<\epsilon_0}$ is the system of ramified analysis up to ϵ_0, which is, roughly speaking, a system with ϵ_0 times iterated elementary comprehension and other suitable axioms[19]. The number ϵ_0, called Cantor's ordinal, is the smallest epsilon number. In mathematics, the collection of epsilon numbers is defined by the property of transfinite numbers of being fixed points of an exponential map:

$$\epsilon = \omega^{\epsilon},$$

in which ω is the smallest transfinite ordinal. As result, they are not reachable from 0 via a finite series of applications of addition, multiplication and exponentiation. The least such ordinal is ϵ_0, which can be viewed as the limit obtained by transfinite recursion of the kind:

$$\epsilon_0 = \omega^{\omega^{\omega^{\cdot^{\cdot^{\cdot}}}}} = sup\{\omega, \omega^{\omega}, \omega^{\omega^{\omega}}, \dots\}.$$

At any rate the relevance of this ordinal relies upon another factor: it is considered the ordinal of PA since PA proves transfinite induction for any ordinal up to ϵ_0. For this reason ϵ_0 might be seen as a natural limit for the number of iteration of the theory ACA of arithmetical comprehension. In spite of this, stronger truth systems have been considered as well.

However, we are interested in the way in which the result has been obtained more than in the result *per se* since the proof-theoretical strength of DT has been established via relative interpretations. Indeed, relative truth-definability is a tool to compare conceptual aspects of theories of truth, but it turns out to be useful in order to establish relevant results even from an instrumentalistic point of view. This seems to confirm the idea of the methodological pluralism in reductions, especially between theories of truth whose intrinsic duplicity allows different kinds of analysis.

[19] For more information on ramified analysis see Feferman [12].

3.3 One more comparison

In the paper *On a Russellian Paradox about Proposition and Truth*, Cantini introduces a formal theory of truth related to Aczel's Frege structures[20] and some common features can be noted between this theory and the theory DT. Furthermore, Feferman in [18] after presenting the axioms of DT observes a similarity between the axioms for D and T and the clauses for proposition and truth in Frege structures and at the same time he stresses several differences we shall see throughout this section. My purpose is to investigate if, and to what extent, the theories PT and DT are comparable, in this discussion the notion of Frege structure will be used as touchstone. In the last part a new version of the theory DT is presented, which has some specific features that make it a 'bridge theory' between the ones considered. Then, the viability of a comparison via interpretation is investigated. Lastly, philosophical points are stressed.

3.3.1 The theory PT

The genesis of the theory PT is characterized by a deep relation between the logical notion of set and the semantical notions of truth and propositions: the starting point are paradoxes concerning sets, in particular a version of the Russell's paradox that involves proposition and truth. Since we are interested in PT as a theory of truth, less attention will be paid to the complex background and the philosophical context, I refer to the quoted paper for a more detailed survey. It suffices to say that PT is built as framework to successfully deal with the Russell's paradox.

PT comprises the axioms of combinatory logic with some further axioms. The language contains the logical operators $\neg$, $\wedge$, $\rightarrow$ and $\forall$. As usual we use the dotted symbols $\underset{\cdot}{\neg}$, $\underset{\cdot}{\wedge}$, $\underset{\cdot}{\rightarrow}$, $\underset{\cdot}{\forall}$ as individual constants representing them. Moreover, there are the predicate symbols $=$, T, P for identity, truth and proposition represented by the individual constants $\underset{\cdot}{=}$, $\underset{\cdot}{T}$, $\underset{\cdot}{P}$. Moreover, a map $\phi \mapsto [\phi]$ can be defined so that it assigns to each formula of the language a term $[\phi]$, called *propositional object*, with the same set of free variables of the formula. Intuitively, this corresponds to the arithmetization in a non-arithmetic contest: it associates a 'name', that is a term, to any formula. Lastly, we define a *propositional function* as a function whose values are propositions, that is to say f is a propositional function if and only if $\forall x(P(fx))$.

Note that below I adopt the infix notation unlike the original paper in which a prefix notation is adopted, namely I write for example $a\underset{\cdot}{\rightarrow}b$ instead of $\underset{\cdot}{\rightarrow}ab$.

Definition 3.3.1. The system PT consists of all axioms of combinatory logic

[20] See Aczel [1].

and the following axioms:

(PT1) $P([\phi]) \wedge (T([\phi]) \leftrightarrow \phi)$ for each $\phi \in \mathrm{AtFml}_{\mathcal{L}_0}$

(PT2) $T(a) \to P(a)$

(PT3) $P(a) \to T([P(a)])$

(PT4) $P([P(a)]) \to P(a)$

(PT5) $P([T(a)]) \leftrightarrow P(a)$

(PT6) $T([T(a)]) \leftrightarrow T(a)$

(PT7) $P(a) \to (\neg T(a) \to T(\underset{\cdot}{\neg} a))$

(PT8) $T(\underset{\cdot}{\neg} a) \to \neg T(a)$

(PT9) $P(\underset{\cdot}{\neg} a) \leftrightarrow P(a)$

(PT10) $P(a) \wedge (T(a) \to P(b)) \to P(a \underset{\cdot}{\to} b)$

(PT11) $P(a \underset{\cdot}{\to} b) \to (T(a) \to P(b))$

(PT12) $P(a \underset{\cdot}{\to} b) \to (T(a) \to T(b) \to T(a \underset{\cdot}{\to} b))$

(PT13) $T(a \underset{\cdot}{\to} b) \to (T(a) \to T(b))$

(PT14) $P(a) \wedge P(b) \leftrightarrow P(a \underset{\cdot}{\wedge} b)$

(PT15) $T(a \underset{\cdot}{\wedge} b) \leftrightarrow T(a) \wedge T(b)$

(PT16) $\forall x P(fx) \leftrightarrow P(\underset{\cdot}{\forall} f)$

(PT17) $T(\underset{\cdot}{\forall} f) \leftrightarrow \forall x (T(fx))$

The axioms are closely related with the *logical schemata* used by Aczel, moreover there are axioms that regulate the relationship between the two collections of propositions and truths, as an instance the axiom (PT2) states that the former contains the latter as subset. In addition, (PT3) guarantees that those claims about P that are true are internally true as well. Then, there is a group of axioms representing the closure conditions for the predicate P that I shall examine later. However, I can anticipate that just like DT a strictly compositional interpretation of the logical operators is chosen. Lastly, the axioms concerning T state how to evaluate the truth of an atomic expression formed by a predicate and of a compound expression built using logical operators. These are very natural axioms for truth since they are nothing but the usual recursive clauses expressing the compositional nature of the notion of truth restricted to P. Moreover, PT proves all T-biconditionals: if ϕ is an arbitrary formula, then

$$\mathsf{PT} \vdash P([\phi]) \to (T([\phi]) \leftrightarrow \phi).$$

Of course just the restricted version of the Tarskian schema can be derived, but I have already supported the naturalness of this choice once this principle is restricted with respect to the range of significance of the truth predicate. These features make PT an interesting theory even adopting a philosophical point of view towards truth theories. A more comprehensive analysis of the axiomatization will be done by comparing it with DT in the following subsection.

This theory is shown to be consistent by providing a model. The ground universe is a combinatory algebra $\mathcal{M}$, whose universe is M. In order to give an interpretation of the two predicate 'to be a proposition' and 'to be a truth', a pair $X = (X_0, X_1)$ of subsets of M is considered, by adopting the same strategy of the Aczel's construction. Among the possible pairs of subsets of M we consider those satisfying the condition $X_1 \subseteq X_0$. Let $\mathcal{F}$ be the family of such pairs. On this family a binary order relation can be defined:

$$X \leq Y \Leftrightarrow X_0 \subseteq Y_0 \wedge \forall a \in X_0 (a \in X_1 \leftrightarrow a \in Y_1),$$

and it can be proved that $\langle \mathcal{F}, \leq \rangle$ is complete partial ordering[21]. Then an operator Γ is defined: $\Gamma(X) = (\Gamma_0(X), \Gamma_1(X))$. $\Gamma_0(X)$ is a collection of those objects satisfying the following formula $A_0(x, X)$:

$$\begin{aligned} \exists u \exists v\ [\ & (x = [u = v]) \vee (x = [Pu] \wedge u \in X_0) \vee \\ & \vee (x = [Tu] \wedge u \in X_0) \vee \\ & \vee (x = (\underset{\cdot}{\neg} u) \wedge u \in X_0) \vee \\ & \vee ((x = (u \underset{\cdot}{\vee} v) \vee x = (u \underset{\cdot}{\wedge} v)) \wedge u \in X_0 \wedge v \in X_0) \vee \\ & \vee (x = (u \underset{\cdot}{\to} v) \wedge u \in X_0 \wedge (u \notin X_1 \vee v \in X_0)) \vee \\ & \vee ((x = \underset{\cdot}{\forall} u \vee x = \underset{\cdot}{\exists} u) \wedge \forall y (uy \in X_0))\ ]. \end{aligned}$$

$\Gamma_1(X)$ is defined by the following formula $A_1(x, X)$:

$$\begin{aligned} \exists u \exists v\ [\ & (x = [u = v] \wedge u = v) \vee \\ & \vee (x = [Pu] \wedge u \in X_0) \vee \\ & \vee (x = [Tu] \wedge u \in X_1) \vee \\ & \vee (x = (\underset{\cdot}{\neg} u) \wedge u \in X_0 \wedge u \notin X_1) \vee \\ & \vee (x = (u \underset{\cdot}{\vee} v) \wedge u \in X_0 \wedge v \in X_0 \wedge (u \in X_1 \vee v \in X_1)) \vee \\ & \vee (x = (u \underset{\cdot}{\wedge} v) \wedge u \in X_0 \wedge v \in X_0 \wedge u \in X_1 \wedge v \in X_1) \vee \\ & \vee (x = (u \underset{\cdot}{\to} v) \wedge u \in X_0 \wedge (u \notin X_1 \vee v \in X_0) \wedge (u \notin X_1 \vee v \in X_1)) \vee \\ & \vee (x = \underset{\cdot}{\forall} u \wedge \forall y (uy \in X_0) \wedge \forall y (uy \in X_1)) \vee \\ & \vee (x = \underset{\cdot}{\exists} u \wedge \forall y (uy \in X_0) \wedge \exists y (uy \in X_1))\ ]. \end{aligned}$$

The operator Γ engenders two classes of objects that, intuitively, satisfy the axioms of the theory: $\Gamma_0(X)$ follows the axioms concerning P and it

[21] That is a partial ordering such that every chain has a least upper bound.

clearly depends on X_0 (the previous candidate for the set of propositions), but in the case of implication it depends on X_1 as well; on the other hand the defining formula of $\Gamma_1(X)$ embodies the recursion clauses of truth. Moreover, it can be observed that $A_1(x, X)$ is built so that objects satisfying it at the same time satisfy $A_0(x, X)$, this means that if $X_1 \subseteq X_0$ then $\Gamma_1(X) \subseteq \Gamma_0(X)$. In other words, if $X \in \mathcal{F}$, then even $\Gamma(X)$ is so. Moreover Γ turns out to be monotone on the structure $\langle \mathcal{F}, \leq \rangle$. This yields to the existence of fixed points, i.e. sets $X \in \mathcal{F}$ such that $X = \Gamma(X)$. Lastly, these fixed points are proved to be models of PT, which is accordingly consistent.

3.3.2 Comparing DT and PT

This presentation reveals immediately some common feature with DT: the choice of a range of significance for the predicate 'to be true' and the restriction of all principles concerning truth to the condition of being in this domain. Now let us have a look at the macroscopic dissimilarities and similarities between the described theory and DT in relation to the basic framework, the predicates, the axiomatizations and the models.

Basic framework

The basic framework represents the first, immediate difference. DT is a truth theory built by following the pattern described in the first chapter: the language of arithmetic is extended by a predicate T for truth and suitable axioms for T are added to the arithmetical ones. In contrast, the framework in which PT is built is the combinatory logic[22].

Let us now explain how this choice affects the thorny issue of the *truth bearers*, i.e. the question of which kind of object should be considered as true or false. For theories like DT the truth bearers are (numbers standing for) sentences, a collection inductively defined starting from the atomic ones up to the compound sentences. But, as said before, sentences are always sentences of a particular language. Accordingly, the axioms reflect the behaviour of the truth predicate with respect to that particular language with its predicates and building rules for terms, formulae and sentences. And if something can proved to be true in the system then it is a sentence of its language. In other words there is a limit imposed by the particular linguistic framework that cannot be transcended. This somewhat threatens the generality of the axioms and of the whole theory.

Adopting the combinatory logic, a more general framework is obtained: objects being true or not true are propositions and no more sentences of a particular language. Sentences are a syntactical category defined by induction on the construction of the formulae, namely a very well-specified

[22] A formal system introduced by Haskell Curry in the 1920s.

collection; whereas, propositions are a collection whose extension is not inductively specified. They only have to meet some non-restrictive closure conditions. Therefore, an axiomatic system like this lends itself well to a good deal of interpretations: a specific language can be from time to time added to the base language in order to obtain applied versions of the theory PT with a specific interpretation. For example PT can be extended with a predicate N for the set of natural numbers, constants for 0 and successor, induction schema for N. Of course even a theory based on the combinatory logic is not framework-independent, but the framework itself it is more general and so less constraining. As a consequence, the value of the proposal of taking the combinatory logic as generalized syntax for truth theories is the chance of a more general interpretation.

Moreover, in a combinatory framework the application of a term to itself is allowed, so the possibility of the self-reference is guaranteed from the structure itself.

Predicates

In the system PT the predicate T isolates among the propositions those which are *truths* and it trivially corresponds to the truth predicate T of DT, of course with the proviso that the object which truth is attributed are different.

What about the other predicates? Let us see to what extent the predicate P, that is the predicate 'to be a proposition', can be related to D, 'to be a determinate and meaningful sentence'. Exactly as for the truth predicates, P and D are properties of different objects: P is applied to objects representing propositions in a combinatory framework, while the extension of D is a set of natural numbers representing sentences of the language of PA with the truth predicate. Apart from that, their intended interpretations are closer than they appear. Feferman holds that the extension of D coincides with the collection of those sentences that are either true or false. Although in the works considered it is not explicitly specified how to understand the intension of the predicate 'to be a proposition', it is widely accepted that being a proposition is something related with the property of being either true or false, or the property of being a truth-value. That is why the axioms of the two theories can be easily compared.

Another point must be stressed: in the system PT the predicate 'to be a proposition' is taken as primitive. This is not the case for D: it is explicitly defined in terms of T. That is why, more correctly, PT ought to be considered a theory of truth and propositions rather than just a truth theory.

Axiomatizations

Focusing on the axiomatic systems it is evident to what extent some essential differences between the axioms of DT and the principles for propositions and truth in Frege structures are balanced by PT.

Let us consider the following remarks:

1. As said in section 3.1.1, the closure conditions for D are assumed to be invertible. However, this is not the case for the predicate 'to be a proposition' used in Aczel [1]. In order to have a perfect correspondence the axioms (DT3)–(DT6) should be weakened by replacing '$\leftrightarrow$' by '$\rightarrow$'. Nevertheless, the axioms of PT imply a strict compositionality: not only the predicate P is closed under the propositional operations and quantifiers but, in the opposite direction, a compound expression is a proposition *only if* all its constituents are so. As said before this choice is reflected in the interpretations of the logical operators, especially disjunction and existential quantifier.

   It is worth focusing again on the axioms concerning implication, whose formulations are non-standard, similarly to Aczel's treatment of '$\rightarrow$' in Frege structures. A standard formulation would be $D(a \underset{\cdot}{\rightarrow} b) \leftrightarrow D(a) \wedge D(b)$ for DT and $P(a \underset{\cdot}{\rightarrow} b) \leftrightarrow P(a) \wedge P(b)$ for PT. However, the chosen formulations can be justified in terms of deduction and reasoning under assumptions: the determinateness (or the property of being a proposition) of the consequent holds under the hypothesis that the antecedent is true. Moreover, even if the antecedent is not true provided that it is determinate (or it is a proposition) then the conditional is determinate (a proposition) whatever the consequent.

2. Although the closure condition for D and P are similar, it must be noted the way in which the theories formulate the condition for an expression of implicative form to be a proposition. The axiom (PT11) does not correspond to the matching direction in the axiom (DT5) and the latter is likely to be underivable in PT.

   In spite of this, if we look at the model we can see that the formula $P(a \underset{\cdot}{\rightarrow} b) \rightarrow P(a) \wedge (T(a) \rightarrow P(b))$ is true in the model. In order to justify the last statement, let us deeper analyse the formulae $A_0(x, X)$ and $A_1(x, X)$. They 'translate' syntactical claims into information to fill the extensions of the sets of truths and propositions. In the set of propositions ($\Gamma_0(X)$) we put all the objects x that satisfy the formula $A_0(x, X)$, namely: all the expressions of equational form, all the expressions of the form $T[u]$, $\underset{\cdot}{\neg} u$, $u \underset{\cdot}{\vee} v$, $u \underset{\cdot}{\wedge} v$ when u and v are already propositions, all the expressions of the form $u \underset{\cdot}{\rightarrow} v$ when u is a proposition and v is a proposition provided that u is a truth and, lastly, all the expressions $\underset{\cdot}{\forall} u$ and $\underset{\cdot}{\exists} u$ when u is a propositional function and uy is a proposition for all objects y. Regarding the axiom

(PT11), there is no trace in the formula $A_0(x, X)$ of a discrepancy with respect to the other clauses: in the inductive process all of them behave like biconditionals. As a consequence, if we considered the formula $P(a \underset{\cdot}{\to} b) \leftrightarrow P(a) \wedge (T(a) \to P(b))$ as axiom instead of (PT10) and (PT11) the resulting theory would be still consistent by the same model. An interesting point would be to find a countermodel, i.e. a model of PT in which $P(a \underset{\cdot}{\to} b) \to P(a) \wedge (T(a) \to P(b))$ is not true.

3. Among the Aczel's principles there are no conditions for atomic propositions like $T\ulcorner\phi\urcorner$, so the axioms (DT2) and (DT8) do not match anything. Unlike Aczel's clauses, PT contains axioms for proposition of the kind 'a is true' just like (DT2) and (DT8), they are, respectively, (PT5) and (PT6).

4. As far as the axioms for T are concerned, an evident correspondence can be found between the second group of axioms of DT, the logical schemata that inductively define the collections of proposition and truth in Aczel and the axioms for T in PT. All the principles concerning truth are restricted so that they hold just for those object that are provable to be propositions or determinate sentences. In any case, the truth predicate, for objects in its range of significance, satisfies the usual recursive defining conditions, namely the standard Tarski conditions.

   However, we shall see in a while that differences of formulations still remain and those represent an obstacle to an interpretation.

5. It does not matter that different groups of connectives are taken as primitive, since the principles concerning the missing ones are easily derivable. For example PT proves:

$$P(a \underset{\cdot}{\vee} b) \leftrightarrow P(a) \wedge P(b),$$

$$P(a) \wedge P(b) \to (T(a \underset{\cdot}{\vee} b) \leftrightarrow T(a) \vee T(b)).$$

Models

In the previous section a standard model for DT has been described. As a reminder, this construction is broadly composed of two steps: first, a 2-valued standard model of PA is expanded to a 3-valued model by using a Krikpe-style construction in which the interpretation of the truth predicate is given by two sets, the extension and the antiextension. The assignment is fixed according the Feferman Logic, i.e. the inner logic of DT. Then, the least fixed point model is converted into a 2-valued model by giving T a classical interpretation again.

The first, trivial difference between this structure and the one built as a model of PT is the ground universe: a model for DT is obtained by expanding a standard model $\mathbb{N}$ of PA with a suitable interpretation of the truth predicate, while the model for PT has an extensional combinatory algebra as base structure. Secondly, as said before, in PT the predicate 'to be a proposition' is a primitive predicate and it requires a suitable interpretation on the semantical level. This must be taken into account in the construction of a model for a theory of truth and propositions. In this regard, an interesting aspect shared by the PT's model and the Frege structures is the *simultaneity* in the generations of the collections of propositions and truths: the clauses are entangled (e.g. in the case of implication) and this requires an inductive definition that generates the propositions and simultaneously gives conditions for their truth. Note that also in the case of a model for DT there are two collections to fill (the extension and antiextension of the truth predicate), however, the inductive step is independent for each of them, that is why a standard monotone inductive definition is enough.

3.3.3 Obstacles to a reduction

Beside a comparison obtained by comparing immediate features, one may wonder whether the theories are comparable by using one of the tools we have seen in the second chapter. Comparing DT and PT it seems that the most suitable tool is *relative interpretability* since the non-logical symbols of DT may have a natural possible definition in PT. In order to show that a theory S is relatively interpretable in T we need a possible definition for each non-logical symbol of $\mathcal{L}_{\mathsf{S}}$ in $\mathcal{L}_{\mathsf{T}}$ and, moreover, we need a primitive recursive translation function mapping the formulae of $\mathcal{L}_{\mathsf{S}}$ in formulae of $\mathcal{L}_{\mathsf{T}}$ with the following requirements: the translation function should preserve the logical structure of the formulae, relativize all the quantifiers of S-theorems and replace any non-logical symbol of $\mathcal{L}_{\mathsf{S}}$ with its possible definition. Then, S is relatively interpretable in T if for each formula of $\mathcal{L}_{\mathsf{S}}$ if S proves it, T proves its translation.

However, interpreting directly DT in PT or *vice versa* is problematic because of the differences in the basic framework and the underlying logic. For this reason, I consider an applicative variant of DT, DT^c, whose quantifiers range over the same unspecified objects of PT.

A bridge theory

I reformulate DT taking as base theory the combinatory logic; this new version of the theory is called 'applicative' since the binary function of application has a central role in such a framework. Terms are generated from variables and individual constants via application and formulae in the usual way from atoms. As for PT I assume the base theory to be extended with a predicate N for the set of natural numbers, constants for 0, successor, predecessor and conditional on N, the induction schema for natural numbers. Moreover, the language contains primitive recursive representations for logical operators and predicates; as usual they are written with the dotted notation. The atoms of the base language, $\mathcal{L}_0$, are of the form $t = s$ or $N(t)$. Accordingly, the axioms of DT should be modified in a suitable way, especially those who explicitly refer to the base theory.

Beside the base theory, another main difference between DT and PT is the treatment of the predicates D and P: the former is a defined predicate while the latter is a primitive one. This affects the model-theoretic part of the theories. The attempt of building a theory for filling the gap between PT and DT can follow two different paths. On the one hand, the basic idea of PT about the role of propositions can be totally accepted. This would imply a formulation of DT in which the predicate D is 'replaced' by the predicate P, to be a proposition, considered as a primitive predicate. The resulting theory would be a theory extremely close to PT. Nevertheless, I believe that this kind of modification would be too artificial and would misrepresent the underlying conception of the theory of determinate truth. To ensure that the resulting theory is just a variant of the source theory any changes should not distort the content of the theory and a simple change of base theory does not. So, in the spirit of DT, even in the applicative variant the domain of the truth predicate retains its property of being explicitly defined. In this case the 'bridge theory' would be just the axiom system of DT with the combinatory logic as base theory. I call this variant DT^c, c standing for *combinatory (logic)*. It is formulated as follows:

Definition 3.3.2. The system DT^c consists of all axioms of combinatory logic and the following axioms:

(DT^{c}1) $D([\phi])$ for each $\phi \in \mathrm{AtFml}_{\mathcal{L}_0}$

(DT^{c}2) $D([T(a)]) \leftrightarrow D(a)$

(DT^{c}3) $D(\underset{.}{\neg} a) \leftrightarrow D(a)$

(DT^{c}4) $D(a \underset{.}{\vee} b) \leftrightarrow D(a) \wedge D(b)$

(DT^{c}5) $D(a \underset{.}{\rightarrow} b) \leftrightarrow D(a) \wedge (T(a) \rightarrow D(b))$

(DT^{c}6) $D(\underset{.}{\forall} f) \leftrightarrow \forall x D(fx)$

(DT^{c}7) $T([\phi]) \leftrightarrow \phi$ for each $\phi \in \mathrm{AtFml}_{\mathcal{L}_0}$

(DT^{c}8) $D(a) \rightarrow (T([T(a)]) \leftrightarrow T(a))$

(DT^{c}9) $D(a) \rightarrow (T(\underset{.}{\neg} a) \leftrightarrow \neg T(a))$

(DT^{c}10)$D(a\underset{.}{\vee}b) \rightarrow (T(a\underset{.}{\vee}b) \leftrightarrow T(a) \vee T(b))$

(DT^{c}11)$D(a\underset{.}{\rightarrow}b) \rightarrow (T(a\underset{.}{\rightarrow}b) \leftrightarrow (T(a) \rightarrow T(b)))$

(DT^{c}12)$D(\underset{.}{\forall}f) \rightarrow T(\underset{.}{\forall}f) \leftrightarrow \forall x(T(fx))$

In DT the axioms (DT1) and (DT7) refer explicitly to the base theory, the former stating that all the atomic sentences of the base theory are determinate and the latter stating how to evaluate the truth of a formula of the base language. The combinatory logic comprises the standard equational logic, namely its atomic sentences have either the form $t = s$ where t and s are terms of the language or $N(t)$, where N is the predicate for natural numbers. Accordingly, the original axioms are translated into (DT^{c}1) and (DT^{c}7). Except from this, the closure conditions for D and T have been left unchanged.

Steps toward an interpretation

Before dealing with the truth-theoretical part of the theories (which is the one we are interested in), I shall rigorously define how a relative interpretation *iota* between generic theories in a combinatory framework is built.

The set of terms is defined inductively as follows: variables and constants are terms and the only terms constructor is the application. In a natural way we assume that:

$$\begin{aligned} \iota(x) &:= x, \\ \iota(c) &:= c, \\ \iota(ts) &:= \iota(t)\iota(s). \end{aligned}$$

Note that since combinators are constants as well, we have for example:

$$\iota(K) := K.$$

In the base language, without the truth predicate, the only descriptive symbols are $=$ and N. This two predicate are translated by themselves, namely:

$$\begin{aligned} \iota(t = s) &:= \iota(t) = \iota(s), \\ \iota(N(t)) &:= N(\iota(t)). \end{aligned}$$

Compound formulae are generated inductively from atoms by propositional operators and quantifiers. The logical structure of the formulae must be preserved, so for all formulae ϕ and ψ of the source theory:

$$\begin{aligned}\iota(\neg\phi) &:= \neg\iota(\phi),\\ \iota(\phi \vee \psi) &:= \iota(\phi) \vee \iota(\psi),\\ \iota(\phi \wedge \psi) &:= \iota(\phi) \wedge \iota(\psi),\\ \iota(\phi \rightarrow \psi) &:= \iota(\phi) \rightarrow \iota(\psi).\end{aligned}$$

As mentioned before, the axioms of the two theories range on the same object. Accordingly, no relativization on quantifiers is needed: $\iota(\forall x\chi) := \forall x\chi$.

In general the dotted symbols are individual constants representing predicates or logical operators, accordingly they are translated by themselves, except from $\underset{\cdot}{T}$, $\underset{\cdot}{P}$ and $\underset{\cdot}{D}$ that being mentioned occurrences of the matching predicate must be uniformly translated with the constant representing the defining formula of the predicates, namely if $\iota(T(x)) := \theta(x)$ then $\iota(\underset{\cdot}{T}(x)) := \underset{\cdot}{\theta}(x)$.

It still remains to establish how the translation function acts on the the truth predicate and the predicate representing its domain. Having done this we shall wonder whether a reduction between the theories DT^c and PT can be established. As we shall see in a while, it seems to me that in both cases there is some trouble. I shall deal with the two questions separately, just focusing my attention on the truth-theoretical part, i.e. the axioms displayed in the definition of the theories, since I assume that in the other respects the functioning of the translation function ι is defined as above.

Is DT^c interpretable in PT?

We need a possible definition for the non-logical symbols T, the truth predicate of DT^c and D, namely respectively two formulae $\theta(x)$ and $\delta(x)$ in $\mathcal{L}_{\mathsf{PT}}$ such that PT proves the following sentences:

$\delta([x = y])$

$\delta([\theta(a)]) \leftrightarrow \delta(a)$

$\delta(\underset{\cdot}{\neg} a) \leftrightarrow \delta(a)$

$\delta(a \underset{\cdot}{\vee} b) \leftrightarrow \delta(a) \wedge \delta(b)$

$\delta(a \underset{\cdot}{\rightarrow} b) \leftrightarrow \delta(a) \wedge (\theta(a) \rightarrow \delta(b))$

$\delta(\underset{\cdot}{\forall} f) \leftrightarrow \forall x\delta(fx)$

$\theta([x = y]) \leftrightarrow x = y$

$\delta(a) \rightarrow (\theta([\theta(a)]) \leftrightarrow \theta(a))$

$$\delta(a) \to (\theta(\underset{\cdot}{\neg} a) \leftrightarrow \neg\theta(a))$$

$$\delta(a \underset{\cdot}{\vee} b) \to (\theta(a \underset{\cdot}{\vee} b) \leftrightarrow \theta(a) \vee \theta(b))$$

$$\delta(a \underset{\cdot}{\to} b) \to (\theta(a \underset{\cdot}{\to} b) \leftrightarrow (\theta(a) \to \theta(b)))$$

$$\delta(\underset{\cdot}{\forall} f) \to \theta(\underset{\cdot}{\forall} f) \leftrightarrow \forall x(\theta(fx)))$$

The most natural and immediate choices for $\theta(x)$ and $\delta(x)$ seem to be respectively, T — the truth predicate of PT— and P. Henceforth, when the context creates ambiguity I shall write T_1 for the truth predicate of DT^c and T_2 for the truth predicate of PT.

However, there is a first, serious, problem about that: D is not a primitive symbol of the language and, rigourously speaking, in a relative interpretation we do not need a possible definition for it. In the theory DT^c, D is defined in terms of T, so if there is a formula $\theta(x)$ of $\mathcal{L}_{\mathsf{PT}}$ which explicitly defines $T(x)$ then as possible definition for D we should simply take the formula $\delta(x) := \theta(x) \vee \theta(\underset{\cdot}{\neg} x)$. Moreover, we are forced to do this because the translation function should *uniformly* substitute all the occurrences of the truth predicate with $\theta(x)$, even the ones 'hidden' in D, as far as $D(x)$ is just an abbreviation for $T(x) \vee T(\underset{\cdot}{\neg} x)$. As a result, if we choose the simplest translation for the truth predicate, namely:

$$\iota(T_1(t)) := \theta(t) := T_2(\iota(t))$$

then the compulsory possible definition for D must be:

$$\iota(D(t)) := \delta(t) := \theta(t) \vee \theta(\underset{\cdot}{\neg} t) := T_2(\iota(t)) \vee T_2(\underset{\cdot}{\neg}\iota(t)).$$

If we look at the definition of relative interpretation[23] we can observe that conditions (i)–(vi) are fulfilled; so the problem whether ι is a relative interpretation of PT in DT^c is simply reduced to the following:

Problem I For all formulae ϕ in $\mathcal{L}_{\mathsf{DT}^c}$, does it hold that:

$$\mathsf{DT}^c \vdash \phi \Rightarrow \mathsf{PT} \vdash \iota(\phi)?$$

An immediate problem related to the definition of D arises: P is a primitive predicate and despite $T_2(a) \vee T_2(\underset{\cdot}{\neg} a) \to P(a)$ holds in PT the reverse it is not entailed by the axioms. Therefore we are not able to use the axioms of PT in the proofs of the translated axioms. The same holds for all the axioms, accordingly ι, if defined as before, cannot be a relative interpretation from DT to PT, unless a prove of $P(a) \to T_2(a) \vee T_2(\underset{\cdot}{\neg} a)$ is found.

It is not excluded that choosing a less obvious definition of T_1 a successful relative interpretation can be built. This remains an open issue. At any rate, I suspect that there would be a further problem because of the differences in the formulation of the axioms (DT^{c}5) and (PT11) mentioned in the point 3. of the comparison between axiomatizations.

[23] See Definition 2.2.6 on page 38.

Is PT interpretable in DT^c?

In this case we do not encounter difficulties in establishing the definitions: since P is a primitive predicate, namely in all respects a descriptive symbol of the language of PT, nothing prevents us to take as a possible definition for it precisely the predicate D, as follows:

$$\iota(T_2(t)) := \theta(t) := T_1(\iota(t)),$$

$$\iota(P(t)) := \theta(t) \vee \theta(\underset{\cdot}{\neg} t) := T_1(\iota(t)) \vee T_1(\underset{\cdot}{\neg}\iota(t)) := D(\iota(t)).$$

In this way, the gap between P and D is, even though artificially, bridged.

Again we wonder whether PT is interpretable in DT^c via ι:

Problem II For all formulae ϕ in $\mathcal{L}_{\mathsf{PT}}$, does it hold that:

$$\mathsf{PT} \vdash \phi \Rightarrow \mathsf{DT}^c \vdash \iota(\phi)?$$

As usual this is verified by showing whether this hold when ϕ is an axiom of PT. While the translations of the axioms stating the closure condition of P are easily provable in DT^c, a first obstacle is the formulation of the truth axioms. For example, the axioms (PT7) and (PT8) are stronger than (DT^{c}9). So, does it hold that $\mathsf{DT}^c \vdash \iota(\mathrm{PT8})$? It seems to me that in DT^c only the corresponding rule can be proved, i.e.:

$$\mathsf{DT}^c \vdash \iota(T_2(\underset{\cdot}{\neg} a)) \Rightarrow \mathsf{DT}^c \vdash \iota(\neg T_2(a)),$$

which is a weaker claim than:

$$\mathsf{DT}^c \vdash \iota(T_2(\underset{\cdot}{\neg} a) \rightarrow \neg T_2(a)).$$

This is the case for axioms (PT6), (PT13), (PT15) and (PT17).

Another immediate obstacle is the fact that DT^c does not contain equivalent axioms for (PT3) and (PT4). Furthermore, their translations, i.e respectively $D(a) \rightarrow T_1([D(a)])$ and $D([D(a)]) \rightarrow D(a)$, are likely to be underivable in DT. However, I suspect they can be added as axioms to DT^c with the proviso that they might be redundant.

To sum up in both cases there are some obstacles to a reduction between these two theories when the reduction itself is carried out by using a natural and immediate interpretation function which translates the descriptive symbols of the source theory preserving their roles. However, as seen before, sometimes interpretation results are obtained by building less trivial and *ad hoc* translation functions. It is not excluded that this can be done for DT and PT as well.

3.3.4 Philosophical remarks

I take this comparison as a further example of the fact that a meta-theoretical analysis between theories of truth can stimulate a purely philosophical debate about the involved thoeries. Even the negative results and the obstacles can be starting points for philosophical observations.

The first remark is related to the basic framework. I have already emphasized that the value of choosing the combinatory logic as general syntax is the greater generality. A further confirmation of this feature derives from a reflection about how the quantifiers are treated in the interpretation. Suppose to interpret directly DT in PT and *vice versa.* I want to stress that for the former an *unrestricted* interpretation is enough while for the latter a relativization of quantifiers is needed. Usually in a relative interpretation the relativization of quantifier is required in order to preserve the provability of the formulae of the source theory in the target theory: reducing S to T, whatever is S-provable about S-objects is translated into something provable in T about the matching T-objects, that are a subset of the domain of T isolated by a relativizing formula. In the axioms of DT the quantifiers range over sentences. How should they be translated in order to be provable in PT? The objects of DT are natural numbers and the set of sentences in DT is suitable subset of this domain, while a combinatory universe is based on an abstract notion of object, whose nature is irrelevant, the central notions being the ones of application and combinator. This shows the elegance of a combinatory setting: it allows a multiplicity of interpretations so that whatever is provable in DT for specific objects, in PT is provable for all the objects. Accordingly, in the translation of the formulae quantifiers shouldn't be restricted. On the other hand, if we wanted to interpret PT in DT then we would restrict the quantifiers and the relativizing formula would be the formula that defines the set of sentences. This shows to what extent a truth theory with a combinatory framework is more general than another theory with the usual arithmetical framework.

Moreover, I claimed that a comparison between truth systems can shed light on the underlying notions of truth. In this case, what do we learn about them? If, in an axiomatic setting, the predicates are defined by providing rules (i.e. axioms) for their behaviour, then it is easy to see that the two theories describe the 'same' notion of truth based on the Russellian idea of the range of significance: all the desirable principles concerning truth (disquotationality, compositionality, completeness, coherence) are preserved and restricted to objects that are in the domain of truth predicate. The value of this choice stands in the fact that it meets the philosophical *desiderata* partially expressed in Leitgeb's criteria and, at the same time, it wards off the risk of inconsistency.

This is what they share. However, the comparison has shown some differences between the theories and, accordingly, some difficulties in the interpretation. I want now to point out that the difference between the treatment of P and D seems to be considerable from a philosophical point of view as well. In many respects, the two predicates can be considered similar given that they share their interpretation and their role with respect to the truth predicate. Nevertheless, a fundamental difference still remains: P is a primitive predicate and this means that the notion of truth which PT relies on is deeply related with the notion of proposition. This is an interesting point since the notion of proposition is anything but marginal and it is philosophically laden, at least as much as the notion of truth.

One may wonder whether it is worth complicating the picture by introducing another inhabitant in a theory of truth. The answer is by no means trivial because, as I shall explain, the choice of taking as a domain of the truth predicate another predicate totally independent from T is, in some respects, more justifiable than the Feferman's one. This is due to a sort of circularity: on the one hand, in the axiomatic system DT the conditions on D are prior to those on T up to the point that each formula $\phi_T(x)$ expressing a principle about truth is subordinate to D: $\forall x(D(x) \rightarrow \phi_T(x))$. On the other hand, the interpretation reflects an opposite trend: the predicate 'to be true' is the first and the only in receiving an extension. In other words, first, in an effective way, we establish which sentences of the language are true and then, since D is defined in terms of T, we use this information in order to give D an extension and establish which sentences are meaningfully determinate. This circularity is avoided once the truth predicate and its domain are taken as separate, like in PT. In this case the two sets simultaneously receive an extension: the inductive process generates the class of propositions and, at the same time, gives conditions for their truth[24].

There is another point related with the issue of circularity: a pretended trivialization of the concept of range of significance for a predicate if explicitly defined in terms of the predicate itself. Compare the following answers to the question: what is the domain of the predicate P?

(i) a set of objects for which it makes sense to say whether they satisfy P or not.

(ii) a set of objects that actually satisfy P or not.

In the first case there must be a choice in the isolation of the set, whereas the second one is a somewhat redundant statement. Let me resort to simpler example: consider the predicate 'to be transitive' instead of the predicate 'to be true', the universe being the set of the 'linguistic expressions' or 'parts

[24] Another route is described by Feferman in [18]: to build a theory of (determinate) meaningfulness followed by a theory of truth. From a semantical point of view this would imply that, in a first stage, the set of sentences that are meaningful and have determinate truth value is isolated and, then, the truth predicate is applied just to terms standing for those sentences.

of speech'. We cannot say whether an article is transitive or not, so it makes sense to isolate a subset of the universe that acts as domain for the predicate. The obvious choice is the sets of verbs: we can meaningfully apply this predicate just to some objects, the verbs, which constitute its domain. This means that $Tr(a)$ is true or false only when a is a verb, otherwise it is not defined. So, if something is transitive or intransitive then it is a verb. On the other hand, the set of verbs contains (all and) only those objects that are transitive or intransitive. By extensionality, the two set coincide: the domain of the predicate 'to be transitive' is the set of verbs and, equally, the set of those object that are transitive or intransitive. However, from an intensional point of view there is a remarkable difference. While the former gives us a defining formula for the domain and, accordingly, a characteristic function (x is in the domain of the predicate if and only if is it a verb), the latter is merely an analytic statement (in a Kantian sense): no further information is given since it is trivially true that each predicate has in its domain all and only the object for which it is true or false that they satisfy the predicate itself. This does not help at all in the isolation of the domain. Nevertheless, even if an identification of (i) and (ii) in the natural language cannot be avoided, this does not hold in a formal language if not deliberately established. What about the domain of the truth predicate? There are two possibilities:

(i) the set of propositions, the set of meaningful and determinate sentences and so on.

(ii) the set of those sentences that are either true or false.

In the first case the set representing the domain is defined by a formula, $D(x)$ or $P(x)$, which is not better specified. In the second, this formula is explicitly defined as $T(x) \vee F(x)$. Although the fact that (ii) $\Rightarrow$ (i) should be accepted, the reverse can be omitted and the circularity avoided. The case of PT shows that isolating a domain by an undefined predicate suffices to avoid paradoxes, without commitments with the definition of this predicate. The problem is that another primitive predicate should be introduced. As said before, the choice is not trivial since one can either stake his all on the truth predicate, but incurring in the difficulties seen before, or resort to other notions transcending the bounds of a truth theory.

All things considered, although the theories are very similar both in the philosophical motivations and in the axiomatizations, the difference in the treatment of P and D represents a remarkable hindrance when one tries to interpret one theory in the other. In particular, if the predicate is a primitive one it can be defined in the other theory assuring that it becomes dependent on the possible definition of the truth predicate. Nevertheless, there is no way to use the undefined predicate as a possible definition for the defined one. Since the choice of taking D as primitive or not affects both the technical formulation and the philosophical assessment of the theory, it might be worth investigating a variant of DT in which D is a primitive predicate. If D and T were taken separately so that $D(a) \nrightarrow T(a) \vee F(a)$, the resulting theory would be a theory extremely close to PT and a model for it would be very similar to the one built by Cantini.

Conclusion

It can be useful to reassert the underlying question of the whole work: a philosophical assessment of the metatheoretical investigation in the domain of axiomatic theories of truth. One of the main advantage in dealing with axiomatic theories is the chance to compare them by using well-known tools and this possibility has been widely explored in the literature. The aim of this work was to submit this practice to a philosophical analysis underlining motivations, results and open issues always bearing in mind throughout the discussion the underlying perspective: an interaction between logical methods and philosophical issues.

The starting point of this work has been a characterization of the axiomatic approach toward truth as a possible solution to semantical paradoxes like the Liar one. The problem of comparing truth theories has been, then, introduced by means of a short analysis of the notion of reduction between theories in general followed by the central issues: what does it happens when truth theories are subject to reduction? Which are, if any, the philosophical side effects? The theoretical analysis has been in turn followed by the introduction of a case study, namely an example of how a theory of truth, DT for *determinate truth*, is compared with other theories. I believe that the study of this theory is prominent for its own sake as well as far as DT is one of the most recent and promising truth theories for at least two reasons: its proof-theory has been investigated with new tools, like truth-definability, with interesting outcomes and, moreover, DT relies upon a natural and coherent philosophical stance towards truth based of the isolation of a range of significance for the truth predicate.

Turning our attention on what came out from this investigation, there are at least two points I hope to have emphasized both from a theoretical point of view and by providing a case study.

First: *whatever approach one adopts towards axiomatic theories of truth, metatheoretical inquiry, i.e. comparing them from various points of view, is an essential tool.* Axiomatic theories of truth can be compared to each other and with other theories, like the base theory or second-order theories of arithmetic. From this investigation, it turns out that certain axiomatic theories of truth are reducible to certain others. A comparison between truth theories can tell us something more about the conception of truth behind them and provide an answer to important questions; for example one can prove the consistency of a theory of truth by showing that it is reducible to another theory which is known to be consistent.

Several important results have been obtained comparing truth theories by their proof-theoretic strength, that is comparing them according to consequences (theorems) they prove. In order to determinate their strength, truth theories have been related in terms of proof-theoretic reducibility to subsystems of second-order arithmetic, whose strength is well-known. Behind this stance there is a specific aim: truth predicate added to a theory can increase its deductive power and, moreover, crucial principles required by the theory can be superseded by truth theoretic axioms. In this way, one can develop comprehensive parts of mathematic in axiomatic theories of truth, showing that the notion of truth can play a role in the foundations of mathematics. In order to pursue this instrumentalistic approach it is enough to consider only the arithmetical, namely truth-free, consequences of truth theories. However, the set of truth-free sentences of a truth theory does not determinate univocally the theory itself; indeed truth theories which embody different view of truth can share the same arithmetic consequences. Therefore, if one adopts a more philosophical approach, preferring to investigate conceptual aspects of truth, then it might be important to consider the entire truth theory including theorems with the truth predicate. To this aim, other methods of reduction must be employed, like different kinds of relative interpretability such as relative truth-definability. From a purely philosophical point of view, reductions between truth systems can tell us something about the compatibility of the underlying notion of truth and contribute to the philosophical debate about it.

Secondly: *the philosophical debate about reduction between formal systems can benefit from this kind of study as well.* That is to say, the behaviour of truth theories once submitted to reduction can shed light on the larger problem of the comparison between theories. There are many different open issues about this theme, in the second chapter we have just pointed out three of them:

- **The dispute about the priority between notions of reducibility.** Various methods to explain inter-theoretical relations have been picked out and, moreover, they are not always equivalent or comparable. The most general and widely used notions of reducibility between theories are *proof-theoretic reduction* and *relative interpretation* (together with their variants and subtypes). As a reminder, we repeat an intuitive definition of these notions: assume that T and S are deductive systems formulated in the languages $\mathcal{L_T}$ and $\mathcal{L_S}$, respectively. Informally, a relative interpretation of S in T is a translation of $\mathcal{L_S}$ in $\mathcal{L_T}$ that preserves the logical structure of the formulae, such that if S proves a formula, T proves its translation. Equally roughly, the system S is proof-theoretically reducible to T if and only if there is an effective method to transform every proof of S into a proof of the same theorem in T, and this is established in a third system or in T itself. The Niebergall-Feferman dispute[25] concerns those notions of reducibility: Niebergall questions that proof-theoretical reducibility is a generally acceptable reducibility concept while, in his opinion, relative interpretability is so and, on the other hand, Feferman argues proof-theoretical reducibility to be the prime candidate for a general relation of reducibility between systems. **Proposal:** I argue in favour of a *methodological pluralism*. Which notion of reducibility is appropriate depends on the purpose of the comparison and the employment of a notion rather than the other might give unexpected and relevant results. In this respect, a sort of 'pragmatism' should be pursued in the choice of tools and methods. Considerations of epistemological and philosophical character ought to be postponed, the priority being the effective use of a wide range of technical instruments. This may apply in general, however, the world of axiomatic theories of truth is a field where philosophy and logic are deeply interwoven and, as a consequence, the necessity of not being an hindrance to each other becomes remarkably urgent.

- **Philosophical relevance of reduction results.** Hofweber raised the issue of the philosophical relevance of reductions. **Proposal:** I argue the feature that makes reduction between truth theory philosophically relevant is a sort of adherence between axiomatizations and their objects. The conception of truth behind a truth theory it is not something separate from the theory itself; it is rather embodied in the axiom system up to the point that these two aspect (axiomatization and stance toward truth) are indivisible. That is why every variation operated on the axioms (extensions, comparisons, translations and so on) can be 'read' from a conceptual point of view. This insight can be useful in general since the criterion of the adherence can be used as touchstone or discriminant even for reductions involving other kind of theories.

- **The relationship between theory reduction and ontological reduction.** This issue is closely related to the previous one: it seems that for Hofweber a theory reduction is philosophical relevant as far as it entails an ontological reduction. **Proposal:** in the field of theories of truth we find proposals about reduction that might be interesting for other fields as well. In particular we mention the Halbach's one of considering ontological commitments as concerning 'assumptions about objects' more than 'objects *per se*'.

[25] See their contributions in *Erkenntnis* 53 (2000).

Since we are dealing with a particular set of theories, namely axiomatic theories of truth, caution is required: one has to take into account within the discussion the peculiarities of the domain of interest. Some working hypothesis might turn out to be appropriate for a specific field, but there is no guarantee they can be ‘exported’ as they are to other fields. I do not argue that this can be done in practice, indeed the viability of this kind of extension should be examined case by case. However, in any case, those hypothesis can constitute a starting point for further considerations.

To sum up, I maintain that reductions can be successfully used in the field of axiomatic theories of truth in order to tackle logical-philosophical issues concerning truth and, moreover, a purely philosophical reflection on reduction and truth may contribute to the debate about reduction *tout court*.

Future development lines may result from two fronts: a logical research and a purely philosophical one. As far as the merely technical part of the work, there are some problems that still remain open. Further investigations might clarify whether those claims are provable or might establish that they cannot be proved — and this would be a result as well. On the philosophical side, I believe a deeper analysis of the philosophical value of this kind of researches is required in order to clarify to what extent a logical research in the field of axiomatic theories of truth make its contribution to the fascinating philosophical issues concerning truth.

Bibliography

[1] P. Aczel. Frege Structures and the Notion of Proposition, Truth and Set. In *The Kleene Symposium*. North-Holland Publishing Company.

[2] R. Batterman. Intertheory relations in physics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.

[3] T. Bolander. Self-Reference. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.

[4] G. S. Boolos and R. C. Jeffrey. Cambridge University press. *Fifth Edition*, title = Computability and Logic, year = 2007, Cambridge.

[5] T. Burge. Semantical Paradox. *The Journal of Philosophy*, 76(4):169–198, 1979.

[6] A. G. Burgess and J. P. Burgess. *Truth*. Princeton University Press, Princeton and Oxford, 2011.

[7] A. Cantini. Notes on Formal Theories of Truth. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 35:97–130, 1989.

[8] A. Cantini. On a Russellian Paradox about Propositions and Truth. In G. Link, editor, *One Hundred Years of Russell's Paradox. Mathematics, Logic and Philosophy*. Walter de Gruyter, 2004.

[9] J. Cat. The unity of science. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.

[10] D. Davidson. The Folly of Trying to Define Truth. In K. S. S. Blackburn, editor, *Truth*. Oxford University Press, 1999.

[11] S. Feferman. Arithmetization of Metamathematics in a General Setting. *Fundamenta Mathematicae*, 49:35–92, 1960.

[12] S. Feferman. Systems of Predicative Analysis. *The Journal of Symbolic Logic*, 29:1–31, 1964.

[13] S. Feferman. Systems of Predicative Analysis II: Representations of Ordinals. *The Journal of Symbolic Logic*, 33:193–220, 1969.

[14] S. Feferman. Toward Useful Type-Free Theories I. *The Journal of Symbolic Logic*, 49(1):75–111, 1984.

[15] S. Feferman. Hilbert's Program Relativized: Proof-Theoretical and Foundational Reductions. *The Journal of Symbolic Logic*, 53(2):364–384, 1988.

[16] S. Feferman. Reflecting on Incompleteness. *The Journal of Symbolic Logic*, 56:1–47, 1991.

[17] S. Feferman. Does Reductive Proof Theory Have a Viable Rationale? *Erkenntnis*, 53:63–96, 2000.

[18] S. Feferman. Axioms for Determinateness and Truth. *The Review of Symbolic Logic*, 1(2):204–217, 2008.

[19] H. Field. Saving the Truth Schema from Paradox. *The Journal of Philosophical Logic*, 31:1–27, 2002.

[20] M. Fischer. Minimal Truth and Interpretability. *The Review of Symbolic Logic*, 2:799–815, 2009.

[21] H. Friedman and M. Sheard. An Axiomatic Approach to Selfreferential Truth. *Annals of Pure and Applied Logic*, 33:1–21, 1987.

[22] K. Fujimoto. Relative Truth Definability of Axiomatic Truth Theories. *The Bulletin Of Symbolic Logic*, 16(3):305–344, 2010.

[23] G. Gentzen. The Consistency of Elementary Number Theory. In M. E. Szabo, editor, *The Collected Papers of Gerhard Gentzen.* North-Holland, 1969.

[24] M. Glanzberg. Truth. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.*

[25] V. Halbach. Axiomatic Theories of Truth. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.*

[26] V. Halbach. Truth and Reduction. *Erkenntnis*, 53:97–126, 2000.

[27] V. Halbach. Reducing Compositional to Disquotational Truth. *The Review of Symbolic Logic*, 2:786–798, 2009.

[28] V. Halbach. *Axiomatic Theories of Truth.* Cambridge University Press, New York, 2011.

[29] W. Hodges. Tarski's Truth Definitions. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.*

[30] T. Hofweber. Proof-theoretic reduction as a philosopher's tool. *Erkenntnis*, 53:127–146, 2000.

[31] L. Horsten. *The Tarskian Turn: Deflationism and Axiomatic Truth.* MIT press, Cambridge, 2010.

[32] S. Kripke. Outline of a Theory of Truth. *The Journal of Philosophy*, 72(19):690–716, 1975.

[33] H. Leitgeb. What Theories of Truth Should be Like (but Cannot be). *Phiosophy Compass*, 2:276–290, 2007.

[34] B. E. McDonald. On Meaningfulness and Truth. *Journal of Philosophical Logic*, 29(5):433–482, 2000.

[35] K.-G. Niebergall. On the Logic of Reducibility: Axioms and Examples. *Erkenntnis*, 53:27–61, 2000.

[36] S. Orey. Relative Interpretations. *Mathematical Logic Quarterly*, 7:146–153, 1961.

[37] C. Parsons. The Liar Paradox. *Journal of Philosophical Logic*, 3(4):381–412, 1974.

[38] G. Priest. What is So Bad About Contradictions? *The Journal of Philosophy*, 95:410–426, 1998.

[39] G. Priest and K. Tanaka. Paraconsistent logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.*

[40] W. N. Reinhardt. Some Remarks on Extending and Interpreting Theories with a Partial Predicate for Truth. *Journal of Philosophical Logic*, 15(2):219–251, 1986.

[41] B. Russell. Mathematical Logic as Based on the Theory of Types. *American Journal of Mathematics*, 30(3):222–262, 1908.

[42] S. Simpson. *Subsystems of Second Order Arithmetic.* Perspectives in Logic. The Pennsylvania State University, 2006.

[43] L. Sklar. Types of Inter-Theoretic Reduction. *The British Journal for the Philosophy of Science*, 18(2):109–124, 1967.

[44] S. Soames. *Understanding Truth.* Oxford University Press, Oxford, 1999.

[45] A. Tarski. The Concept of Truth in Formalized Languages. In Tarski (1956), pages 152–278, 1935.

[46] A. Tarski. *Logic, Semantics, Mathematics.* Clarendon Press, Oxford, 1956.

[47] A. Tarski. The Semantic Conception of Truth and the Foundation of Semantics. In K. S. S. Blackburn, editor, *Truth.* Oxford University Press, 1999.

[48] A. Tarski, A. Mostowski, and R. M. Robinson. *Undecidable Theories.* North-Holland, Amsterdam, 1953.

[49] A. Visser. Categories of Theories and Interpretations. In A. Enayat, I. Kalantari, and M. Moniri, editors, *Proceedings of the Workshop and Conference on Logic, Algebra, and Arithmetic*, pages 284–341, 2004.