Truth, Deflationism, and the Ontology of Expressions: An Axiomatic Study

Thesis submitted for the degree of Doctor of Philosophy in Philosophy

CARLO NICOLAI

TRINITY TERM 2013
SOMERVILLE COLLEGE
UNIVERSITY OF OXFORD

Truth, Deflationism and the Ontology of Expressions: An Axiomatic Study

Thesis submitted for the degree of Doctor of Philosophy in Philosophy

Carlo Nicolai, Somerville College, University of Oxford,

Trinity Term 2013

Word Count: 61000ca[1]

Abstract:

Philosophical enquiry on the notion of truth has traditionally involved the identification of a class of objects to which truth is ascribed. At the same time, formal investigations are often required when the notion of truth is at issue: semantic paradoxes force in fact philosophers to shape their arguments in a precise way. Objects of truth, in formal context, are always reduced to other, more manageable objects that mimic their structural properties such as numbers or sets. This form of reduction renders the distinction between linguistic or syntactic objects, to which truth is usually applied, and their mathematical counterparts opaque, at least from the point of view of the theory of truth. In informal metatheoretic discussion, in fact, they are clearly different entities.

In this thesis we focus on an alternative way of constructing axiomatic theories of truth in which syntactic objects and mathematical objects belong to different universes. A brief introduction tries to situate the proposed theories in the context of different investigations on axiomatic truth. Chapter 2 is devoted to the discussion of historical and more theoretical

[1]The word count cannot be precise as the softwares available consider very long formulas as one single word. It is likely that, using different criteria, the number of words outnumbers the one reported.

motivations behind the proposed alternative. Chapter 3 will present the syntactic *κοινή* spoken by our theories. Morphological categories of the object language and logical concepts concerning the object theory will be formalised in a recent axiomatisation of hereditarily finite sets. In Chapter 4 we finally introduce the theories of truth with 'disentangled' syntax and examine some of their consequences. We briefly focus on disquotational truth, then consider compositional axioms for truth. Chapter 5 investigates a possible application of the setting just introduced: a realisation of the all-present interaction, in metamathematical practice, between informal metatheoretic claims and their (suitably chosen) coded counterparts. In the final chapter, after a brief characterisation of the key doctrines of the delflationary conception of truth, we evaluate the impact that the theories of truth studied in this work can have on the debate on the so-called conservativeness argument, which tries to match the alleged insubstantiality of the notion of truth, advocated by deflationists, with the deductive power of deflationary acceptable theories of truth.

Acknowledgements

This thesis could not have been written without the generous help of others.

I would like to thank the Arts & Humanities Research Council, the Faculty of Philosophy of the Oxford University and the Centro Universitario Cattolico for the financial support received during the nine terms in which this thesis was written.

Thanks to the audiences in Munich, Oxford, Milano, Bristol, Amsterdam, Princeton and Buenos Aires. Special thanks go to Sergio Galvan, Alessandro Giordani and Ciro de Florio for their helpful comments to my (two) talks in Milan.

Thanks to Richard Heck and Jeffrey Ketland, who gave me access to unpublished notes that significantly contributed to the developments of the ideas presented in this work. Jeff Ketland also attended and commented my talks in Munich, Amsterdam and Buenos Aires: again his comments had a remarkable impact on my work.

Thanks to Kentaro Fujimoto, Graham Leigh, Lorenzo Rossi and James Studd for their encouragement and support. James Studd shared with me most of the time in which this thesis was written in our office in the Philosophy Faculty: his help and encouragement have been essential. Kentaro Fujimoto had the patience of punctually answering to my trivial questions either in person or by email and encouraging me at early stages when I needed it most. Special thanks go to Graham Leigh: our collaboration put the basis of many of the arguments carried out in this work. Although I reduced to the minimum the overlaps between this work and our joint paper, his influence is undeniable.

Daniel Isaacson read many parts of this work: his objections and comments have surely made this thesis better.

Most of all, I would like to thank my supervisor Volker Halbach: he inspired this line of research and followed the development of the project with great patience.

Oxford, September 2013

Contents

1 Introduction

The present work is concerned with the justification, the formal investigation and the philosophical evaluation of a nonstandard way of constructing axiomatic theories of truth. The framework is essentially based upon the possibility of cutting the ties between the object theory, usually a formalisation of a portion of mathematics, and the theory formalising the basic structural properties of the objects to which truth is ascribed. It essentially differs from the usual way of constructing axiomatic theories of truth:[1] in the customary setting, in fact, syntactic objects are identified with objects in the domain of discourse of the object theory.

Ouline. This work is divided in six chapters including the present introduction.

Chapter 2 will be devoted, in its first part, to the the study of the genesis and of the reasons behind the inclusion of the theory of syntax into the object theory by focusing on the relationships between Tarski's project of allowing a consistent use of the notion of truth in formalised languages and the concurrent studies on metalogical properties of formal systems carried out by Gödel. In the second part, we will consider some peculiar consequences that stem from the customary way of axiomatising the truth predicate over schematically axiomatised base theories.

Chapter 3 we introduce the syntactic *κοινή* spoken by our theories: that is, we present the theory that will deal with the 'syntactic' component of our framework. Morphological

[1]For a thorough survey of axiomatic theories of truth, we refer to [Halbach 2011] and [Horsten 2011].

categories of the object language and logical concepts concerning the object theory will be formalised in an axiomatisation of hereditarily finite sets that has recently received much attention. The choice of this particular theory will be motivated in the following way: a detailed investigation of theories of truth with a 'disentangled' syntax calls for a specification of a theory dealing with the 'morphology' of the object theory, and we opted for a theory which naturally captures the way in which the syntax of formal languages is usually (informally) understood.

In Chapter 4. we finally introduce the theories of truth with 'disentangled' syntax and examine some of their consequences. We briefly focus on disquotational truth, then consider compositional axioms for truth. We then compare the resulting theories to the object theory, focusing on conservativity and relative interpretability. Various choices of the object theory are taken into account. We conclude the chapter by considering a strengthening of our theory of truth obtained by expanding axiom schemata of the object theory,[2] if present, to the entire language of our theory of truth.

Chapter 5 investigates a possible application of the setting just introduced. The extension of the theory of truth with suitable axioms determining a bijection between objects of truth and objects in the sense of the object theory can be taken to offer a realisation of the all-present interaction, in metamathematical practice, between informal metatheoretic claims and their (suitably chosen) coded counterparts.

In the final chapter, after a brief characterisation of the key doctrines of the delflationary conception of truth, we evaluate the impact that the theories of truth studied in this work can have on the debate on the so-called conservativeness argument, which tries to match the alleged insubstantiality of the notion of truth, advocated by deflationists, with the deductive power of deflationary acceptable theories of truth.

In the remaining parts of this Introduction we outline some of the distinctive features of this alternative construction and provide some general remarks that are intended to

[2] For the notion of axiom schemata that will be assumed throughout this work, see Chapter 2.

supplement the content of the five remaining chapters. References will be kept to the minimum in these introductory remarks: bibliographical data are provided in full detail in the next chapters.

A Tarskian way. The framework that we are going to investigate in this work appears to be a direct axiomatisation of Tarski's orginal picture of the metatheory as presented in his game-changing paper *Der Wahrheitsbegriff in den formalisierten Sprachen* ([Tarski 1956a]), in which he provided a definition of the notion of true sentence for suitable formalised languages. The importance of Tarski's paper can hardly be overestimated:

> Tarski's article *The concept of truth in formalized languages* (1935) [A/N: my emphasis] is a momentous achievement. In it we find the same kind of radical shift as we find in Galilei and Newton with respect to the concept of force. In a sense, Tarski's move can be seen as an emancipation of truth theory from traditional philosophy. ([Horsten 2011, p. 15])

According to Horsten, Tarski's work provided a new way of approaching the problem of truth: from the the Socratic $\tau\iota\ \varepsilon\sigma\tau\iota$ question, which called for an account of the 'essence' of truth, Tarski suggested to focus on the way in which truth 'functions'.

The contribution that axiomatic theories of truth provide to philosophy, if any, goes in this direction: some principles describing the 'functioning' of the truth predicate are added to a base theory and the expressive strength and deductive power determined by the axioms for truth is evaluated with respect to the underlying base theory. The resulting theories, however, do not have to be intended as mere reconstructions of how the truth predicate is used in natural languages. On the contrary, they can be viewed as attempts to examine consequences and metatheoretic properties of systems based on some conception of the role of truth that could possibly contribute to an emendation of natural language.

It is known that Tarski provided a *definition* in the metatheory of the notion of true sentence for the language of the object theory that does not contain the truth predicate itself, whereas we claimed that we are going to capture the spirit of the Tarskian construction by means of an axiomatic framework. It is natural to wonder whether we are misinterpreting the nature of Tarski's enterprise.

One aspect of the Tarskian picture that we are clearly not misunderstanding is the presence, in the axiom system compounding the metatheory, of an independent axiom system devoted to the formalisation of the syntax of the object theory, whose language is disjoint from the language of the object theory itself. Besides the peculiarities of the Tarskian setting—he was using an second-order, categorical version of the syntax—it seems that this similarity is undeniable.

However, the objection was based upon the role of the notion of truth, and this is the main point that has to be addressed. We do not intend to defend the claim that the theories treated in this work are *faithful* to Tarski's original motivations. For us it would be sufficient to show that they may be regarded as highlighting some distinctive features, proper of Tarski's original construction, that are hidden in the usual axiomatisations.

As [Halbach 2011] also notices, Tarski starts his project with the recognition of some *principles* of syntactic nature that were intended to be recovered from his definition of truth in the metatheory: in particular he identified as unavoidable consequences of any satisfactory account of truth all instances of the schema

$$\text{`}p\text{' is a true sentence if and only if } p \tag{1.1}$$

where p is a metavariable for sentences of the object language and 'p' is name of that sentence given in the metalanguage. The entire project seems thus to be driven by some postulates of syntactic nature.

In Section 5 of the *Warheitsbegriff*, however, Tarski directly considers the possibility of an axiomatisation of truth based upon those principles, and dismisses this axiomatisation because of its deductive weakness.[3] In particular, the system obtained from the theory containing the syntax theory and the object theory by expanding its language with the truth predicate and by adding to it, as new axioms, the formalisation of all instances of the schema (1.1)—which resembles closely the system TBD[O] introduced Ch. 4—cannot prove some important general claims such as 'the general principle of contradiction', which states that it is not the case that a sentence of the language of the object theory and its negation are both in the extension of the truth predicate.

But we do not need to confine ourselves to weak theories, proving only instances of (1.1). Since Tarski's paper it has become quite clear, among philosophical and mathematical logicians, that a formal theory of truth should describe the way in which the truth predicate interacts with logical constants. If we extend the theory sketched by Tarski above with the so-called *compositional axioms* for truth, which directly derive from the recursive clause of the Tarskian definition, we obtain a theory which is able to prove the general principle of contradiction and many other generalisations of syntactic nature. This theory finds a sufficiently accurate characterisation in the theory CTD[O]—also introduced in Chapter 4. CTD[O] meets Tarski's adequacy criteria, and it looks at least compatible with the epoch-making picture drawn by Tarski in the thirties.

In conclusion, if an axiomatisation of the truth predicate based on Tarski's compositional clause, constructed in the usual way, can be taken as a legacy of Tarski's work to some extent, we will see that the axiomatisation of truth introduced in this work seems even closer to the spirit of the Tarskian enterprise.

An essentially typed approach. In a nutshell, in *typed* (axiomatic) theories of truth the axioms governing the truth predicate characterise only its application to sentences not containing the truth predicate. In *type-free* (axiomatic) theories, on the other hand, the truth predicate applies provably to sentences involving the truth predicate.

[3] Cfr. [Tarski 1956a, p. 257] and [Halbach 2011, Ch. I.3] and Ch. 4 of the present work for a discussion of Tarski's adequacy criteria.

In a more inclusive sense, Tarski's construction is intrinsically typed, in the sense that the object language, although included in the metalanguage, does not contain the truth predicate defined in the metalanguage. Given what we just said in the previous section, it is thus no surprise that our axiomatisations will reflect the nature of Tarski's definition.

We might also discriminate between typed and type-free axiomatic theories of truth by looking at their intended models: a type-free theory would be identified by the presence of sentences containing the truth predicate in the extension of the truth predicate, as defined in an intended model of the theory. If one does not want to resort to models, it would be necessary to formulate general criteria of demarcation between typed and type-free theories of purely syntactic form. [Halbach 2011, Ch. 10] makes an attempt in this direction: roughly, a theory of truth is *typed* if and only if (i) the truth predicate provably applies only to sentences not containing the truth predicate, and (ii) the axioms for truth do not stipulate anything about the truth of sentences containing the truth predicate. These clauses can be made precise only if one makes specific assumptions about the base language, but it is not hard to see how to proceed for the usual choices.

The theories presented in Chapter 4 are formulated in a three-sorted language $\mathcal{L}_{\mathsf{TD}}$: quantifiers of the first sort range over the domain of discourse of the object theory; quantifiers of the second sort over 'syntactic' objects; and quantifiers of the third sort range over sequences of variable assignments which connect the syntactic and the mathematical realm. Consistently with the presentation that we will consider in Ch. 4, we denote with $x, y, z, \ldots$, with $i, j, k, l \ldots$ and with $a, b, c, \ldots$ respectively 'object-theoretic' variables, 'syntactic' variables and 'sequential' variables. The notion of truth will be captured by means of a binary satisfaction predicate $\mathsf{Sat}(a, k)$, applying to an assignment for variables of the first-sort and to a term representing a formula of the language of the object theory whose denotation lives in the 'syntactic', disjoint realm. For the details of the theories, we refer to Ch. 4; for the sake of this introduction, we just needed to fix some notation to consider the status of our theories in relation to the demarcation between typed and type-free axiomatic theories of truth.

Theories of truth with 'disentangled' syntax satisfy Halbach's criteria for typed-theories of truth in virtue of purely syntactic features of their language: the satisfaction predicate in fact can only be applied to terms whose variables belonging to a certain type and satisfying some further syntactic constraints, such as ranging over certain syntactic categories of the language of the object theory. Other formulations of the theories, which avoid typing of variables, can still mimic the purely syntactic restrictions proper of the many-sorted setting. Theories of truth with disentangled syntax seems to display a fundamentally typed nature. *Is this a downside of our theories?*

In some sense it is. If one is interested in evaluating the contribution that axiomatic theories of truth provide to the formulation of a semantics for natural language, in fact, typing might be considered not to be a reasonable choice, as there is no typing in natural language. However, the impact that axiomatic theories of truth can have on the semantics of natural languages is not clear.[4] Axiomatic theories of truth in fact, type-free theories included, are constructed taking sentences to be mathematical objects, and important features of natural language such as the presence of indexicals is completely ignored.

There are, however, more serious drawbacks: in particular, type-free theories boost our capability of proving general claims. This can be extracted from the proof-theoretic analysis of type-free theories such as Kripke-Feferman, Friedman-Sheard or PUTB[5]—just to mention theories formulated in classical logic—which are notoriously stronger than the classical axiomatisations of Tarskian truth. In addition, philosophy is populated with claims containing truth-iterations; therefore philosophical argumentation would be seriously undermined if an axiomatic theory of typed truth was employed to formalise philosophical claims involving the truth predicate.

By looking at our specific purposes, however, it seems that there is no need to worry about

[4] Cfr. [Halbach 2011, Ch. 11].
[5] We still refer to [Halbach 2011] and [Horsten 2011].

these general issues concerning typing. In the first place, Tarski's solution to the paradoxes, based upon the object language/metalanguage distinction, is still the starting point of contemporary formal semantics and model theory. Mathematicians have assimilated Tarski's solution, and the notion of truth-in-a-model plays an important role in current mathematics, also outside what is commonly identified as logic.[6]

In Chapter 5 we will provide a formal setting that is intended to capture some relevant aspects of metamathematical reasoning: in particular, the close relationships existing between informal metamathematical claims their counterparts in the language of the object theory. An paradigmatic example is represented by the following situation:

> Suppose you have a object theory O formalising some portion of mathematics. To the acceptance of O correspond beliefs on the soundness (call it $\mathsf{Sound}^{\mathsf{O}}$) and consistency (call it $\mathsf{Cons}^{\mathsf{O}}$) of O. The informal metatheoretic claim $\mathsf{Sound}^{\mathsf{O}}$ involves a truth predicate which is not in the language of O. $\mathsf{Sound}^{\mathsf{O}}$ obviously entails $\mathsf{Cons}^{\mathsf{O}}$. In metamathematical practice, for a suitable choice of O, we require $\mathsf{Sound}^{\mathsf{O}}$ to be captured by the global reflection principle for O,[7] and $\mathsf{Cons}^{\mathsf{O}}$ to correspond to a canonical consistency statement for O, both formalised in the language of O.

If one tries to formalise the simple, intrinsically typed reasoning sketched in the quotation, our framework seems to be appropriate. Chapter 5 will deal with the case in which O is a theory in a pure mathematical language. In a forthcoming work, we will consider more complicated cases.

Besides the characterisation of the interactions between informal and formal metamathematical claims, we will also consider the contribution that theories of truth with 'disentangled' syntax can give to the debate about the so-called conservativeness argument against deflationism. The debate rotates around the compatibility of the deflationary claim that

[6] Algebraic geometry, for instance.
[7] Cfr. §6.2.

truth in an insubstantial property that does not play any role in philosophical and scientific argumentation on the one hand, and of the nonconservativity—over the base theory—of any satisfactory axiomatisation of truth on the other. For the sake of an analysis of this debate, there is no need to consider type-free theories. The debate has in fact been carried out in the typed environment: only typed, axiomatic theories of truth obtained by adding to the base theory axioms capturing the disquotational and compositional properties of the truth predicate have been taken into account.

A satisfactory account of the notion of truth, it has been argued, should be sufficient to prove that all the theorems of the object theory are true. For natural choices of the object theory, this is only possible when compositional axioms for truth are present and axiom schemata of the object theory are extended to the truth predicate: we will introduce this theory under the label of CT when the base theory is Peano Arithmetic. However, the most influential type-free theories, still constructed over Peano Arithmetic, all contain CT as subtheory. If one wants to focus on the conservativeness argument then, it seems that there are no strong motivations to take into account theories stronger than CT.

The theories of truth that we introduce in this work are intended to mimic the construction of a typed, compositional theory of truth in the new framework given by the disentanglement of the syntax. Also for the second application that we want to make of our framework, therefore, it seems that its typed nature does not represent a huge limitation.

2 Objects of Truth and Metamathematics

2.1 Objects of Truth

Philosophers disagree on the entities to which truth has to be ascribed. Those entities are known as bearers of truth. Although beliefs, judgements and even ideas have been considered as primary truth bearers, it seems that contemporary debates contemplate sentence tokens and propositions as the only reasonable candidates.

Experience tells us that we attribute truth to certain sounds or inscriptions. By example, we may have:

(2.1) Nicole said 'Lilies are flowers'. That is true.

Quotation marks are in fact linguistic devices used to construct names for particular sentences. In other words, we often attribute truth to *sentence tokens*. As [Burgess&Burgess 2011] notice, however, the notion of sentence token is ambiguous. A *phonological* or an *orthographic* token is (respectively) a sound or an inscription to which no meaning is associated; whereas a *semantic* token can only be produced by someone who speaks or writes meaningfully. For someone who believes that primary objects of truth are sentence tokens, semantic tokens are usually taken to be the genuine bearers of truth. Phonological and orthographic tokens can in fact display context-dependent ambiguities related, for instance, to the presence of indexicals.

Propositions are even more often taken as primary objects of truth by philosophers.[1] Instead of explaining what they are, which is a matter of millenary controversy, we limit ourselves to showing how they manifest themselves in language. They are expressed by what is commonly called a *that-clause*:

(2.2) Nicole said that Lilies are flowers.

The linguistic construct involving the occurrence of 'that' and the sentence 'Lilies are flowers' is a that-clause and it is usually taken to denote the proposition expressed by the sentence 'Lilies are flowers'. Propositions share with semantic (sentence) tokens the resistance to features of context and to indexicals and also respond in a nice way to translations. Propositions, however, are still fairly mysterious entities and propositionalists disagree on how to understand them.

Since we are mainly concerned with an analysis of how the objects to which truth is ascribed are conceived in formal contexts, it seems sufficient for our purposes to accept the status quo in which sentence tokens or propositions are fighting for the position of primary bearers of truth and to regard theories of sentence types as the means by which it is possible to talk about our primary truth bearers in formal contexts.

This approach seems in fact to be independent from our preferences regarding sentence tokens or propositions. In particular, if we believe that propositions are primary truth bearers, we can read á la Quine the expression 'the sentence x is true' as 'the sentence x expresses a true proposition' (see [Quine 1986]). Therefore a theory of sentence types, via this translations, can also talk about the truth of propositions. At the same time, one can argue that theories of sentence types are useful tools to talk about utterances or concrete

[1] For instance, [Horwich 1998] and [Soames 2010].

strings of symbols which are emitted or shaped according to the rules of formation of sentences proper of one's grammar.

Sentence types, however, seem to be a useful reduction available in extensional contexts only; they are also affected by forms of context-sensitivity. The sentence

(2.3) I'm proud of Paolo's performance

is true only when I am uttering it after good game played by the volleyball player Paolo, but it might not be true after a severe defeat if I utter it speaking to a third person in Paolo's presence. As a consequence, the explication of the truth of a proposition in terms of sentential truth is not always available, especially when intensional contexts are taken into account. As we will only be concerned with extensional contexts, however, sentence types will play the role of the objects to which truth is ascribed in this work.

In many formal accounts of truth, and in all theories considered below, truth is conceived as a predicate, whose extension is directly or, more often in a derivative sense, the class of objects of truth. The base theory—or equivalently object theory—of our theory of truth is in fact usually taken to be able to interpret a theory of sentence types. In this way, the base theory is sufficiently rich to prove some basic facts concerning structural properties of the objects of truth mimicking those of sentences.

If one is a *sui-generist*, and she is worried about ontological reductions, what we have just sketched does not sound promising; nonetheless these repeated reductions are customarily seen as harmless simplifications. To be able to appeal to Ockham's razor, in fact, one should be able to show that this proliferation of objects is indeed *praeter necessitatem*.[2]

Variants have indeed been considered: if truth is not understood as a predicate one can dispense with a set of objects of truth. In particular, if truth is described in terms of special sorts of quantifiers, as in [Williams, M. 1986] and, above all, in [Grover, et alii 1975], there

[2] Cfr. Chapter 4.

is no need to specify a class of entities falling into the extension of the truth predicate. Alternatively, one might consider propositions to share the same structure with sentences and consider a theory of sentence types as a theory of propositions.

Giving up the usual assumption that sentence types are, at least in a derivative sense, the objects to which the truth predicate is applied seems to lead, however, to bigger problems than the ontological worries sketched above. If on the one hand prosentential interpretations of truth rely on quite strong—and objectionable—philosophical views,[3] on the other propositions seem to be extralinguistic objects and the assumption that they share the same structure as sentences seems to presuppose the existence of a grammar for propositions as there is for sentences of natural or formal languages.

If one sticks with the sententialist approach, as we are, the picture might be paraphrased in the following way: focusing on a theory of numbers or sets that interprets the theory of sentence types in question—our primary (or secondary already) truth bearers—in any model of the arithmetical or set theory at stake it is possible to construct an internal model which satisfies the axioms of our theory of sentence types. A set of numbers, or a set of (set-)codes, is then taken to be the class of our objects of truth, and properties that are ascribable to sentences in the form of syntactic truths are replicated in the chosen theory and proved for the class of derivative truth bearers so obtained.

The theory of numbers or sets employed in the process just sketched, however, might not be the object theory of our theory of truth yet. It is usually assumed in fact that the object theory contains the theory of numbers or sets employed to capture the properties of sentence types. More precisely, independently of its strength, it could be the case that the object theory is judged too clumsy to talk about syntactic objects and their properties in an effective way, so that it is regarded as more convenient to merely require the more agile theory of the objects of truth to be interpretable in it (see for instance, [Smorynski 1977]).

[3] [Halbach 2001b], for instance, observes that the prosentential theory of truth, by denying that truth is a predicate, seems to deny the existence of any predicate.

From this fact, however, it is a rather short step to plainly assume the theory of the objects of truth to be a subtheory of the object theory.

We try to clarify this intricate picture: primary truth bearers are first reduced to sentence types, that is to types of strings of symbols of a countable alphabet; later, they are identified with numbers or sets—carrying in mind the isomorphism between concatenation structures and natural numbers.[4] Furthermore, a model of this theory of derivative bearers of truth is always available in any model of our object theory, and this justifies the 'reduction' of objects of truth to a particular subset of the domain of the object theory.

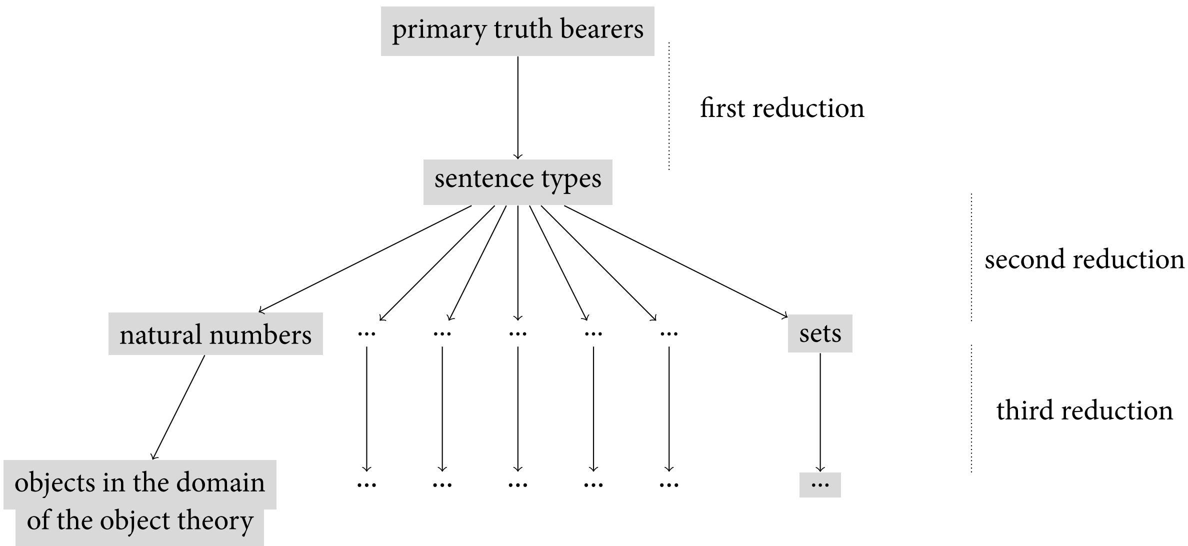

Figure 2.1: Truth bearers in formal contexts.

Later in this chapter we shall argue that the inclusion of the theory of truth bearers into the base theory—a theory formalising some portion of mathematics or science or even philosophy—might determine in some critical cases that patterns of reasoning concerning the objects of truth are confused with patterns of reasoning belonging to the subject-matter

[4]More on this point in the next chapter.

proper of the object theory. Before that, however, we will see that this identification, which is plainly assumed in the usual constructions, is not present in the founding work on formal theories of truth, Tarski's *Wahrheitsbegriff* ([Tarski 1956a]).

2.2 Tarski

It is well-known that Tarski's project of providing a scientifically acceptable characterisation of the notion of 'true sentence' for a wide range of formalised languages relies on the distinction between object-language and metalanguage.[5] It also seems to be folklore that this distinction stems from what is nowadays known as Tarski's Theorem, which states that a theory that is sufficiently strong to develop the *morphology* of its own language cannot define *sine contradictione* the set of true sentences for this language.[6] More precisely:

> Theorem I. (α) In whatever way the symbol '*Tr*' denoting a class of expressions, is defined in the metatheory, it will be possible to derive from it the negation of one of the sentences which were described in the condition (α) of the Convention T;
> (β) assuming that the class of provable sentences of the metatheory is consistent, it is impossible to construct an adequate definition of truth in the sense of convention T on the basis of the metatheory. ([Tarski 1956a, §5, p.247])

It is not necessary to go very far to find the source of this misconception:[7] Tarski himself gave, retrospectively, a fairly inclusive characterisation of his entire project. Some years later in fact he writes:

[5] Since Tarski considers 'languages' as deductive systems, we will employ, for the rest of this section, the expressions object language/object theory and metalanguage/metatheory interchangeably.
[6] Morphology is just the name that Tarski employs to refer to the metatheory itself (see [Tarski 1956a, p.251].
[7] Besides Tarski's quote below, one can also mention this, more modern, passage of [Shapiro 1998, p.494]:

> We know, from Tarski's theorem, that, if a language $\mathcal{L}$ can express a small amount of arithmetic, then no consistent theory in $\mathcal{L}$ can express a truth predicated for $\mathcal{L}$. Kurt Gödel and Alfred Tarski both discovered that one can simulate self-reference in a language of arithmetic.

> Since we have agreed not to employ semantically closed languages, we have to use two different languages in discussing the problem of the definition of truth and, more generally, any problem in the field of semantics. ([Tarski 1944, p. 349])

But what is a semantically closed language? Tarski's understanding seems to consider any language,[8] either formalised or a portion of scientific discourse, featuring names for each of its expressions and containing a truth predicate; moreover, this language would have to be capable of deriving all instances of the unrestricted T-schema.[9] The reason for ruling out semantically closed languages lies obviously in the fact that the antinomy of the Liar can be replicated in them.

We shall argue shortly that motivations behind Tarski's proposal is independent from Theorem I above. The only inconsistency related to semantically closed languages that Tarski was aware of, at least at the time in which his work was sent to press (1933), was the antinomy of the Liar for natural language to which section 1 of his essay is devoted. At any rate, we shall *not* argue that Tarski's theorem is not a sufficient motivation for the entire project. In fact it is the best possible motivation. But the *positive* part of Tarski's construction does not depend on Theorem I, which requires the possibility of interpreting the system of the metatheory in the object theory.

2.2.1 The Problem of Priority and the Generality of the Metatheory.

Recent historical reconstruction have supported the claim that Tarski's theorem does not lie among the original motivations at the basis of Tarski's semantic approach. In the Postscript to the German edition of the *Wahrheitsbegriff*, Tarski explores the relations of 'dependence or independence' occurring between his work and Gödel's.[10] Tarski emphasises the independence of his methods especially in connection to (i) the formulation of the Convention

[8] A theory, in modern terminology.
[9] *Ibid. p. 348*.
[10] Cfr. [Tarski 1956a, p. 277].

T; (ii) the definition of truth for the object language in a sufficiently rich metalanguage; (iii) semantic consistency proofs; (iv) the tripartite structure of the metatheory; (v) the possibility of interpreting the metasystem in arithmetic. On the other hand, it is also made clear by Tarski that the result concerning the undefinability of truth for languages of infinite order—Theorem I, §5, p. 247 in [Tarski 1956a]—was only a later addition:

> I take this opportunity of mentioning that Th. I and the sketch of its proof was only added to the present work after it had already gone to press. At the time the work was presented at the Warsaw Society of Sciences (21 March 1931), Gödel's article—so far as I know—had not yet appeared. In this place therefore I had originally expressed, instead of positive results, only certain suppositions in the same direction, which were based partly on my own investigations and partly on the short report,[11] [...] which had been published some months previously. [...] Moreover, the abstract character of the methods used by Gödel renders the validity of his result independent to a high degree of the specific peculiarities of the science investigated. ([Tarski 1956a, pp. 247-8, footnote 1])

The suppositions mentioned in the quotation amount to the possibility of interpreting the metatheory in arithmetic and an example of a consistent but ω-inconsistent theory.[12]

Tarski's acknowledgement does not leave much room for interpretation, at least concerning the paternity of the theorem. [Gómez-Torrente 2004], however, has argued that Tarski anticipated Gödel in another sense: the formulation and the proof of part (α) Theorem I, in fact, from which part (β) follows, is actually a purely syntactic argument that holds for any predicate other than the truth predicate Tr mentioned in the theorem. In other words, it is a proof of the *fixed-point theorem*, which in its general form was attributed by Gödel to Carnap (proved in [Carnap 1934]) in the Princeton lectures that will be mentioned in a moment. It is not clear whether Tarski appreciated this point at the time, but his proof surely

[11] Here he is referring to a report by Gödel called 'Einige metamathematische Resultate über Entscheidungsdefinitheit und Widerspruchsfreiheit', appeared in the Gazette of the Academy of Sciences of Vienna in 1930.

[12] Which was obtained by Tarski already in 1927 ([Tarski 1956b, footnote 2]).

generalises the construction presented in [Gödel 1931]. For further details concerning the historical reconstruction, we refer to Gómez-Torrente's paper.

Much more controversial, it seems, is the exegesis of Gödel's connections with the undefinability of truth. On the other hand, since Tarski's theorem implies the existence of undecidable sentences, it is reasonable to suspect that Gödel was aware of the undefinability result before Tarski. As reported by [Feferman 1998, p.158], some years after the publication of his paper, Gödel attributed to the undefinability of truth a primary importance in the route towards the incompleteness theorems; the existence of undecidable sentences, according to Gödel, was a consequence of the fact that a 'complete epistemological description of a language A cannot be given in the same language A'.[13]

According to [Wang 1981]—which contains an outline of Gödel's intellectual life in the period under consideration—after his doctoral thesis, Gödel seriously attempted to provide a finitary consistency proof for analysis. The strategy chosen by Gödel was to assume that the consistency of arithmetic could be proved in a finite portion of number theory, and to interpret analysis into arithmetic so to obtain a relative consistency proof of the former. Gödel's idea was to interpret variables ranging over sets of natural numbers as arithmetically definable sets. As Feferman explains, this strategy is equivalent to consider an enumeration of formulas $\varphi_i(x)$ with $i \in \{1, 2, 3, \dots\}$ with only the variable displayed free, collapse sets of natural numbers into $\mathbb{N}$ and define

$$n \in m :\leftrightarrow \text{`}\varphi_m(n)\text{'} \text{ is true.} \tag{2.4}$$

(2.4) requires a general notion of truth for sentences of number theory, besides a suitable machinery to name them. This opens the possibility of replicating the already known semantic antinomies in the new framework.

Gödel's breakthrough was then to show how to construct self-referential statements in

[13]See the introduction of [Von Neumann 1966, pp. 55—56].

the pure arithmetical language. With the help of this method, an analog of the Liar sentence in natural language could be obtained in the arithmetical system considered, and it then followed that the notion of truth for sentences of its language could not be defined by a formula of the language of number theory. Since Gödel's method enables one to construct a formula of the language of arithmetic which defines the set of (Gödel numbers of) theorems of number theory, and all provable formulas are also true, then there must be a formula that is true but not provable.

As Gödel remarked in his letter to Zermelo, the argument for the existence of undecidable sentences just sketched, although perfectly acceptable by his (Gödel's) standards, being of nonconstructive character, could be regarded as suspicious from an intuitionistic point of view. This seems to emphasise how sensitive Gödel was to the intellectual environment in which the metamathematical problems related to his results first arose.

As Feferman points out, the considerations above seem to indicate that the notion of arithmetic truth as a definite concept was essential for Gödel to obtain his theorems. Both Feferman and Murawski ([Murawski 1998]) report the following passage of a letter written by Gödel to Hao Wang, which is quite instructive for the clarification of the matter:

> I may add that my objectivist conception of mathematics and metamathematics in general, and of transfinite reasoning in particular, was fundamental also to my other work in logic. How indeed could one think of expressing metamathematics in the mathematical systems themselves, if the latter are considered to consist of meaningless symbols which acquire some substitute of meaning only through metamathematics...it should be noted that the heuristic principle of my construction of undecidable number theoretical propositions in the formal systems of mathematics is the highly transfinite concept of 'objective mathematical truth' as opposed to that of 'demonstrability' (cf. M. Davis, The Undecidable, New York 1965, p. 64 where I explain the heuristic argument by which I arrive at the incompleteness results), with which it was generally confused before my own and Tarski's work.

Nonetheless the concept of arithmetic truth seems to play a quite marginal role in

[Gödel 1931]. There is also no mention of the adjective 'wahr' in the paper, which is instead replaced by 'richtig' there, and still the latter is employed only the introductory section when he outlines the strategy of the proof of his first theorem.[14] On the other hand the term 'true' appeared in the lectures Gödel gave at the Institute of Advanced Studies in 1934, collected by Kleene and Rosser and reproduced in [Davis 1965]. It is plausible to attribute this absence to the extensive influence that Hilbert's foundational standpoint had on the mathematical community. Gödel himself in fact, in the letter to Wang mentioned above, notices how if one takes the notion of demonstrability within a formal system as an 'analysis' of the concept of truth—as formalists do—, then it is not possible to appreciate the difference between the two notions.

A direct explanation can be found in a reply by Gödel to a letter of Yossef Balas, a student at the University of Northern Iowa, and reported in [Wang 1987, pp. 84-85]. There Gödel observes how the 'philosophical prejudices' of the period led to consider the notion of mathematical truth suspicious—if not meaningless—when opposed to provability in a formal system. [Feferman 1998] thus concludes that Gödel's presentation of his work in terms of finitary methods was intended to render the results contained in his paper appealing also to people that were skeptical about non-finitary methods. Moreover, his effort to shape the results according to the formalist's taste also led him to what is nowadays known as the second incompleteness theorem.

A related question is whether or not Gödel, like Tarski, actually appreciated the importance of a set theoretic definition of truth for arbitrary first-order languages; what has been said above in fact seems to represent a definite answer to the role that the notion of arithmetic truth and its undefinability had in the development of the ideas contained in [Gödel 1931], but there seems to be no hint to a general account of truth-in-a-structure for any first-order language in his work.

There seem to be no striking evidences in favour of a positive answer to this question:

[14]Cfr. [Murawski 1998, p.157].

first of all in his account of the completeness of the first-order calculus the notion of validity is understood informally and not in terms of a satisfaction relation; moreover, in footnote 48a of [Gödel 1931] Gödel describes the possibility of deciding the undecidable propositions of a language by resorting to languages of higher types. This can be carried out by defining a satisfaction relation in a system of set theory T between the language of a system S and the structure $(V_\alpha, \in)$ for some $\alpha \in \mathsf{Ord}$. However, Gödel's explanation of the suggested decision method was never completed as the second part of his paper never appeared.

[Murawski 1998], on the other hand, gives an affirmative answer to the question; according to him Gödel was already acquainted with the general set theoretic method required for the definition of truth. Gödel writes in fact to Carnap in 1932:[15]

> On the ground of these thoughts I will provide in the second part of my work a definition of 'true' and I presume that things cannot be tackled otherwise and that it is possible not to grasp the higher *Induktion Kalkül* semantically but syntactically.

The expression 'the second part of my work', according to Murawski, refers to the joint project commissioned by Springer to Heyting and Gödel of providing a survey of current investigations in mathematical logic, in which a 'syntactic' definition of truth in set theory was supposed to be described. The project, unfortunately, was never completed due to Gödel's health problems.

At any rate, it is clear that Tarski had mainly the set theoretic definition of truth as his main concern, and the pars destruens of the entire project represented by the undefinability results played a secondary role in the development of the ideas presented in the *Wahrheitsbegriff*. The possibility of talking directly of the notion of mathematical truth is also related to the

[15] My translation of

> Ich werde auf Grund dieses Gedankens im II. Teil meiner Arbeit eine Definition f"ur 'wahre' geben und ich bin der Meinung, da"s sich die Sache anders nicht machen l"a"st und da"s man den h"oheren Induktionenkalk"ul nicht semantisch [d.h. damals syntaktisch!] auffassen kann.

cultural environment of the Lvov-Warsaw school, which was more liberal towards non-finitary mathematics than the Vienna Circle, in which Hilbert's ideas were more influential.[16]

It is thus safe to conclude that Tarski did not initiate his project motivated by the impossibility of providing a definition of the notion of true sentences for formal, semantically closed languages. The possibility of interpreting the 'metatheory' into the object theory does not play an essential part in the genesis and in the development of his project. The 'syntactic' or 'morphological' component of the metatheory, in Tarski's original project, was devised as disjoint from the object theory.

In the Poscript to [Tarski 1956a], however, Tarski states that he became aware of the possibility of interpreting the 'morphology' of the object theory into the object theory itself independently from Gödel's work. But when? Was the distinction object language/metalanguage already developed at that time?

The presence of a disjoint theory of syntax in Tarski's original construction has something to do also with the dimension of generality proper of Tarski's initial construction. That the definition devised by Tarski was intended to be applicable to a wide range of choices of the object theory is evident by considering the following passage of the *Wahrheitsbegriff*, which also highlights the interest that Tarski had in weak object theories:

> In contrast to natural languages, the formalised languages do not have the universality which was discussed at the end of the preceding section. In particular, most of these languages possess no terms belonging to the theory of language, i.e. no expressions which denote signs and expressions of the same or another language or which describe the structural connexions between them (such expressions I call—for lack of a better name—*structural-descriptive*). For this reason, when we investigate the language of a formalised deductive science, we must always distinguish clearly between the language *about* which we speak and the language *in* which we speak, as well as between the science which is the object of our investigation and the science in which the investigation is carried out. The names of the expressions of the first language, and of the

[16] See [Wolenski 1989] for more details.

> relations between them, belong to the second language, called the *metalanguage* (which may contain the first as a part). The description of these expressions, the definition of the complicated concepts, especially those connected with the construction of a deductive theory (like the concept of consequence, of provable sentence, possibly true sentence), the determination of the properties of these concepts, is the task of the second theory which we shall call the *metatheory*. ([Tarski 1956a, §2, p.167])

The passage gives a partial answer to the questions that we left open and reinforces our point: the object language/metalanguage distinction seems to be motivated by technical convenience, and not by a strong result such as the impossibility of a definition of the notion of true sentence for universal languages.

Moreover, the passage clearly indicates that Tarski's method for isolating the class of true sentences of the object language has to be applicable also to theories which are not sufficiently rich to formalise syntactic and logical notions concerning the object theory itself. Since the metatheory, as presented in the *Wahrheitsbegriff*, includes or interprets the object theory, the 'richness' considered in the quotation above actually lies in the possibility, for the object theory, to interpret the 'theory of language' forming part of the metatheory. This, in turn, amounts to the possibility of interpreting number theory into the object theory.

To put things in a more modern terminology, we can also say that the 'richness' of the object theory consists in the possibility of interpreting a sufficiently strong fragment of arithmetic in it. But the Calculus of Classes, which plays the role of the object theory in a significant part of Tarski's work, was shown to enjoy a form of elimination of quantifiers and to be mutually interpretable with first-order monadic predicate logic by Thoralf Skolem in [Skolem 1919]. Now a suitable fragment of arithmetic which is strong enough to interpret a theory of concatenation with full induction as the one introduced by Tarski in [Tarski 1956a, p. 173] will surely extend Robisons' Q.[17] But if Q were interpretable in the Calculus of Classes, the latter would be undecidable, which is not the case by Skolem's result and the decidability

[17] For the axioms of Q, see [Hájek&Pudlák 1993, p. 28].

of the monadic calculus. It thus seems to be clear, in the end, that Tarski considered the applicability his method to be independent from the possibility of developing the syntax for the object theory in the object theory itself.

A final question still remains open: if Tarski did not initiate his program inspired by what is now known as Tarski's Theorem, why did he became interested in the notion of truth at all? There are at least two possible answers: on the one hand, following [Betti 2008], Tarski seemed to have inherited the project of a rehabilitation of 'metamathematical' concepts from unpublished work by Ajdukeiwicz. On the other, according to [Hodges 2008], Tarski intended to pursue Leśniewski's philosophical project of 'intuitionistic formalism', according to which the contents of the mind of the proving mathematician should have been laid out explicitly, semantical notions included.

We can summarise our short voyage in the last century as follows: Tarski's goal of isolating the class of true sentences of a great variety of languages does not depend on what is known as Tarski's theorem. This represents a later addition that a posteriori justified in an elegant and powerful way the initial project, and in particular the distinction between object and metalanguage, but that also hid an important dimension of the Tarskian initial construction, that is its wide applicability.

2.3 Objects of Truth and Schematic Reasoning

As it is well-known, the payoff for the 'reduction' of the metatheory to the object theory is mathematically very fruitful. However, having Tarski's picture of the metatheory in mind, the 'collapse' of the morphology of the object language into the object theory still seems to strike as a simplification for certain purposes. In the present section we will provide some evidences in favour of this claim by examining some well-known axiomatisations of Tarskian truth constructed over object theories containing axiom schemata.

2.3.1 Theories and Schemata

Let us start by introducing some terminology. A theory is as a set of sentences closed under logical consequence. We call a theory T *axiomatisable* if and only if there is a recursive set S of sentences such that T is just the set of sentences true in all models of S. We say that T is *finitely axiomatisable* if the set S in question is finite. An *axiomatisation* of a theory T is just a specific set S. As a preliminary characterisation, let us call a theory T *schematically axiomatisable* if, besides the finitely or infinitely many sentences of its language $\mathcal{L}$ in the set S of axioms of T, S also features a linguistic recipe to generate infinitely many sentences of $\mathcal{L}$, its instances.

Particularly important for our discussion will be the examination of how axiomatic theories of truth are constructed over some schematically axiomatised object theories. In nuce, it has been recently argued that *every time that schemata are expanded to contain the truth-predicate in truth-theoretic constructions, a confusion between the syntactic and object-theoretic components of the theory of truth is determined*. In the present section we analyse this claim and contribute to the discussion.

We can define a *schema* as a linguistic pattern compounded by two main ingredients: a syntactic string of elements of the alphabet of the language under consideration together with blanks and/or of placeholders; and a rule, a recipe that enables one to fill the blanks or substitute the placeholders with genuine elements of the alphabet of the language (or suitable extension of it) to generate a possibly infinite list of well-formed expressions. Schemata are thus expression types and its instances are not the corresponding tokes, but just expression types whose tokens are strings of signs of the object-language.[18]

When theories formalising some portion of mathematics are at issue, there are at least two understandings of the notion of an axiom schema:[19] roughly, in the first the schema in considered as a rule, in the second, as an infinite set of formulas. As an example, let us

[18] See [Corcoran 2012] for the above-mentioned definitions and further insights.
[19] Cfr. [Feferman 1991].

consider Peano Arithmetic: we can on the one hand think of PA as specified by a triple $(\mathcal{L}_A, \mathsf{Ax}_{\mathsf{PA}}, \mathsf{Sub\text{-}P})$, where $\mathsf{Ax}_{\mathsf{PA}}$ is the set of basic axioms of PA and Sub-P is a rule that allows one to go from the 'mixed' string

$$P(0) \wedge \forall x(P(x) \to P(Sx)) \to \forall x P(x), \tag{2.5}$$

where x in an object-theoretic variable and P is a new free predicate variable, to strings of the form

$$\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(Sx)) \to \forall x \varphi(x), \tag{2.6}$$

where φ ranges over formulas of $\mathcal{L}$ and $\varphi(t)$ replaces every occurrence of $P(t)$ in (2.5); or on the other hand we can resort to the second notion of a schema and see PA as specified by the pair $(\mathcal{L}_{\mathsf{A}}, \mathsf{Ax}^*_{\mathsf{PA}},)$ where $\mathsf{Ax}^*_{\mathsf{PA}} = \mathsf{Ax}_{\mathsf{PA}} \cup \{\text{all instances of (2.6)}\}$.

The difference is exclusively conceptual, as from the mathematical point of view, the axiomatisation of PA together with Sub-P is conservative over the axiomatisation of PA given by all its instances,[20] but it becomes relevant when the language of the theory at issue, for instance the language of arithmetic $\mathcal{L}_{\mathsf{A}}$, is extended to contain further predicates. The second formulation of PA, in fact, seems to be relative to the language of arithmetic, whereas the first formulation, given the presence of Sub-P, is suitable for the application—in the sense of [Feferman 1991]—of the schematic PA to arbitrary extensions of its language. This conceptual difference has some implications in the very possibility of arbitrarily expanding the schemata of the object theory at any extension of the object language and of formulating some metatheoretic questions concerning the acceptance of the schematic theory in question, such as the determination of its reflecting closure.

In the following, we will not be directly occupied with the question concerning the

[20] *Ibid.* p. 7.

possibility of arbitrarily expanding schemata of the object theory. For our purposes, it is sufficient to embrace the second characterisation of a schema presented above. We will then say that a theory is *schematically axiomatisable* if it has an axiomatisation involving all instances of a schema $\Phi(X)$, where X is a free predicate variable and Φ is a linguistic template relative to the language at issue. In studying particular axiomatisations of the truth-predicate over finitely and schematically axiomatisable theories, we will rather (implicitly) talk about systems and not theories, that is specific sets of axioms and rules, so by saying that the object theory is *finitely (schematically) axiomatised* we will have in mind a particular axiomatisation of it.

2.3.2 Classical Tarskian Truth and Soundness Claims

Let us start with an arbitrary schematically axiomatised (first-order) object theory O in a first order language $\mathcal{L}_{\mathsf{O}}$ extending the language of arithmetic. For simplicity, we assume that O contains $I\Sigma_1$. Therefore, the usual syntactic and logical operations on $\mathcal{L}_{\mathsf{O}}$ are taken to be naturally represented in O by means of a suitable coding of finite sequences.[21] In particular, expressions of $\mathcal{L}_{\mathsf{O}}$ are just identified with numbers, and $I\Sigma_1$ proves the existence of the Δ_1-set Expr, represented in $I\Sigma_1$ by the formula Ex̣pr(x), which is the least set containing all atomic expressions of $\mathcal{L}_{\mathsf{O}}$ and closed under the application of the operations associated with the formation of terms and formulas. To each expression $e \in$ Expr is thus associated an $\mathcal{L}_{\mathsf{O}}$-term $\ulcorner e \urcorner$ such that $I\Sigma_1 \vdash$ Ex̣pr$(\ulcorner e \urcorner)$. Still in $I\Sigma_1$ we define Δ_1 predicates Vạr, Tẹrm, Ctẹrm, AtF̣ml, Fṃl and Sẹnt representing in $I\Sigma_1$ the sets of variables, terms, closed terms, atomic formulas, formulas and sentences of $\mathcal{L}_{\mathsf{O}}$. Also there we find constants for the non logical and logical constants of $\mathcal{L}_{\mathsf{O}}$ so that, for instance, the constants ọ and Ṣ represent 0 and the successor function S—that is the numbers identified with them—in $I\Sigma_1$, and ¬̣$\ulcorner \varphi \urcorner$ expresses in $I\Sigma_1$ the Δ_1 operation of prefixing the negation symbol to the formula φ. Quantification over codes

[21] For a detailed account of the syntax of a first-order language $\mathcal{L}_{\mathsf{O}}$ in $I\Sigma_1$, we refer to [Hájek&Pudlák 1993, pp. 52ff].

is explained in the following way: $Qv\ulcorner\varphi\urcorner(\ldots$ stands for $Qx(\underset{\cdot}{\mathsf{Fml}}(x) \to \ldots)$, and similarly for quantification over codes of terms.

A total Δ_1 function $\underset{\cdot}{\mathsf{num}}(x)$, which assigns to each number its numeral, is definable in $I\Sigma_1$ and abbreviated with $\dot{x}$; also, $I\Sigma_1$ proves, by formalising its course-of value recursion, the existence of a total Δ_1 substitution function $\underset{\cdot}{\mathsf{sub}}(x, y, z)$—also written $x(z/y)$ when x codes a formula of $\mathcal{L}_{\mathsf{O}}$—which formalises the clauses for substitution of free variables for arbitrary terms in formulas and terms of $\mathcal{L}_{\mathsf{O}}$.[22] We will have, in particular,

$$\underset{\cdot}{\mathsf{sub}}(\ulcorner\varphi(x)\urcorner, \ulcorner x\urcorner, \underset{\cdot}{\mathsf{num}}(x)) = \ulcorner\varphi(\dot{x})\urcorner$$

and, more in general

$$\underset{\cdot}{\mathsf{sub}}(\ulcorner\varphi(x)\urcorner, \ulcorner x\urcorner, \ulcorner t\urcorner) = \ulcorner\varphi(t)\urcorner.$$

Quite relevant for our purposes is the Δ_1 evaluation function $\underset{\cdot}{\mathsf{Val}}(x, y)$—provably Δ_1 and total in $I\Sigma_1$—which intuitively applies to the term of $I\Sigma_1$ associated with term of $\mathcal{L}_{\mathsf{O}}$ and to a variable assignment for its free variable and such that for all terms $t(x_1, \ldots, x_n)$ of $\mathcal{L}_{\mathsf{O}}$, with at most the free variables displayed, we can have

$$I\Sigma_1 \vdash \underset{\cdot}{\mathsf{Val}}(\ulcorner t\urcorner(\underset{\cdot}{\mathsf{sub}}(\ulcorner t\urcorner, \ulcorner x_1\urcorner, \dot{x}_1), \ldots, \underset{\cdot}{\mathsf{sub}}(\ulcorner t\urcorner, \ulcorner x_n\urcorner, \dot{x}_n), \langle\rangle) = t(x_1, \ldots, x_n), \tag{2.7}$$

where $\langle\rangle$ denotes in $I\Sigma_1$ the empty sequence.[23] Therefore we can have

$$I\Sigma_1 \vdash \underset{\cdot}{\mathsf{Val}}(\ulcorner t\urcorner) = t \quad \text{for all closed terms } t. \tag{2.8}$$

As in [Halbach 2011], for the sake of readability, the value of the closed term represented in $I\Sigma_1$ by x will be denoted by x°.

Let us now expand the language $\mathcal{L}_{\mathsf{O}}$ of O with a unary truth predicate T. Let O^T be

[22] Cfr. [Hájek&Pudlák 1993, p. 54].
[23] *Ibid.* p. 55.

simply O formulated in the language $\mathcal{L}_T = \mathcal{L}_\mathsf{O} \cup \{T\}$. O^T is clearly conservative over O, as T is nothing more than an uninterpreted predicate symbol in O^T.

We consider the following schemata:

$$\mathsf{Tb} :\leftrightarrow T\ulcorner\sigma\urcorner \leftrightarrow \sigma \quad \text{for any sentence } \sigma \text{ of } \mathcal{L}_\mathsf{O};$$

$$\mathsf{UTb} :\leftrightarrow \forall x(\mathsf{Ct\underset{\cdot}{e}rm}(x) \rightarrow (T\ulcorner\varphi\urcorner(x/\ulcorner v\urcorner) \leftrightarrow \varphi(x^\circ)))$$

for every formula $\varphi(v)$ with exactly the v free.

The theories TB and UTB are then obtained from O^T, in the case in which O is PA, by enlarging the set of its axioms with Tb and UTb respectively.[24] By a result due to Tarski himself, TB and UTB are conservative extensions of PA.[25] Essentially, in any proof in TB and UTB the truth predicate can be replaced by a partial truth predicate definable in PA. Tarski also regarded TB (or UTB) as not adequate theories of truth as important general claims, such as the general principle of contradiction, are not provable in it. Moreover, also the option of adding such general statements independent from TB (UTB) is regarded by Tarski as a not viable option given the arbitrariness of the resulting axiomatisations.[26]

$\mathsf{CT}[\mathsf{O}]$ will result instead from the addition of the clauses of Tarski's definition,[27] turned into axioms, to O^T. The axioms of $\mathsf{CT}[\mathsf{O}]$, assuming that $\mathcal{L}_\mathsf{O}$ has $\neg, \rightarrow$ and $\forall$ as logical constants, are displayed in Table 2.1.

The subsystem $\mathsf{CT}[\mathsf{O}]\!\upharpoonright$ of $\mathsf{CT}[\mathsf{O}]$ is obtained from $\mathsf{CT}[\mathsf{O}]$ by blocking the possibility of the occurrence of the truth predicate into the schemata of O. When O is PA, we will simply write CT and $\mathsf{CT}\!\upharpoonright$ respectively. $\mathsf{CT}\!\upharpoonright$ can be shown to be conservative over it.[28] On the contrary, CT is not conservative over PA, as CT can formalise the standard *semantic consistency proof* for PA. That is to say, we can prove in CT the global reflection principle

[24] In the present and later sections and chapters, we will follow the notation of [Halbach 2011].
[25] For a proof, see [Halbach 2011, p. 55ff].
[26] Cfr. [Tarski 1956a, pp. 257ff].
[27] Cfr. [Tarski 1956a, p. 153].
[28] The first proof is due to [Kotlarsky et alii 1981]. A more modern proof is contained in [Enayat&Visser 2013].

$$
\begin{array}{ll}
\mathsf{CT}R & \forall x \forall y \big(\underset{\cdot}{\mathsf{Cterm}}(x) \wedge \underset{\cdot}{\mathsf{Cterm}}(y) \to (Tx \underset{\cdot}{=} y \leftrightarrow x^{\circ} = y^{\circ})\big) \\
\mathsf{CT}\neg & \forall x \big(\underset{\cdot}{\mathsf{Sent}}_{\mathcal{L}_{\mathsf{O}}}(x) \to (T \underset{\cdot}{\neg} x \leftrightarrow \neg Tx)\big) \\
\mathsf{CT}\wedge & \forall x \forall y \big(\underset{\cdot}{\mathsf{Sent}}_{\mathcal{L}_{\mathsf{O}}}(x \underset{\cdot}{\to} y) \to (Tx \underset{\cdot}{\to} y \leftrightarrow (Tx \to Ty))\big) \\
\mathsf{CT}\forall & \forall x \forall v \big(\underset{\cdot}{\mathsf{Sent}}_{\mathcal{L}_{\mathsf{O}}}(\underset{\cdot}{\forall} v x) \to (T \underset{\cdot}{\forall} v x \leftrightarrow \forall y (\underset{\cdot}{\mathsf{Cterm}}(y) \to Tx(y/v)))\big)
\end{array}
$$

Table 2.1: Axioms of $\mathsf{CT}[\mathsf{O}]$.

for PA. More precisely, we refer to the global reflection principle for PA as the claim

$$\mathsf{GRP}_{\mathsf{PA}} :\leftrightarrow \forall x (\underset{\cdot}{\mathsf{Sent}}_{\mathcal{L}_{\mathsf{PA}}}(x) \wedge \underset{\cdot}{\mathsf{Bew}}_{\mathsf{PA}}(x) \to Tx),$$

where $\underset{\cdot}{\mathsf{Bew}}_{\mathsf{PA}}(x)$ is, provably in $I\Sigma_1$, a canonical provability predicate for PA. In particular, we can safely assume the set of axioms of PA to be Δ_1; there is thus a Δ_1 formula $\underset{\cdot}{\mathsf{Prf}}_{\mathsf{PA}}(x, y)$ which formalises the relation 'x is a proof of the formula y' in $I\Sigma_1$. We then set

(2.9) $\underset{\cdot}{\mathsf{Bew}}_{\mathsf{PA}}(x) :\leftrightarrow \exists y (\underset{\cdot}{\mathsf{Prf}}_{\mathsf{PA}}(y, x))$

(2.10) $\underset{\cdot}{\mathsf{Con}}_{\mathsf{PA}} :\leftrightarrow \forall x \forall y \forall z (\underset{\cdot}{\mathsf{Sent}}_{\mathcal{L}_{\mathsf{PA}}}(x) \to \neg \underset{\cdot}{\mathsf{Prf}}_{\mathsf{PA}}(y, x) \vee \neg \underset{\cdot}{\mathsf{Prf}}_{\mathsf{PA}}(z, \underset{\cdot}{\neg} x))$

For the provability predicate just defined, $I\Sigma_1$ proves the so-called *derivability-conditions*.[29] $\mathsf{GRP}_{\mathsf{PA}}$ thus amounts to the explicit assertion of soundness of PA, which is not expressible in PA itself due to Tarski's theorem.

Finally we have:

Fact 2.1. $\mathsf{CT} \vdash \mathsf{GRP}_{\mathsf{PA}}$.

As an immediate consequence of $\mathsf{GRP}_{\mathsf{PA}}$, $\underset{\cdot}{\mathsf{Con}}_{\mathsf{PA}}$ is provable in CT, which yields the non conservativity over PA.

[29] Cfr. [Hájek&Pudlák 1993, Lemma 2.17, p. 163]. Hájek&Pudlák call them 'provability conditions'.

The proof of of Fact 2.1 has been taken by [Heck 2009] and [Halbach 2011] as a paradigmatic example of where the ambivalent nature of the schemata of the theory of truth appears: in CT there are two possible uses of the extended axiom schema of induction, one *syntactic* and one *mathematical*. We thus briefly recall the main steps of the proof, especially because some interesting details have been overlooked in the philosophical discussions based upon it.[30] We defer a discussion of Heck and Halbach's claim to the next section.

The strategy is based upon a straightforward induction on the length of the derivation in PA: it is first shown that all axioms of PA are true; that the rules of inference preserve truth, to conclude that all theorems of PA are true.[31] Already in $\mathsf{CT}\!\restriction$ we can prove the schema UTb for all formulas $\varphi(v_1, \dots, v_n)$ of $\mathcal{L}_{PA}$.[32] From this the truth of all the basic axioms of PA immediately follows. It remains to be shown the truth of all (closed) instances of induction which, have we have seen, are taken to be part of the axiomatisation of PA at issue. To obtain the claim, it suffices to apply compositional axioms from Table 2.1 to the following instance of the schema of induction of CT, with $\mathsf{V\underset{\cdot}{a}r}(v)$,

$$Tx(\underset{\cdot}{0}/v) \wedge \forall y(Tx(y/v) \to Tx(\dot{S}y/v)) \to \forall y Tx(y/v) \tag{2.11}$$

Let us now define in $I\Sigma_1$ a Δ_1 predicate $\mathsf{P\underset{\cdot}{r}v_{PA}}(x)$, expressing that x is a finite sequence and that all entries of x are either logical axioms, or axioms of PA, or obtained from earlier elements of the sequence by application of the rules of inference of the chosen calculus in which PA is formulated. Moreover, we let $\mathsf{e\underset{\cdot}{n}d}(x)$ be the Δ_1 function, provably in $I\Sigma_1$, which outputs the universal closure of the last entry of the sequence x.

The induction on the length of the derivation in PA informally outlined above is captured

[30] See for instance the discussion of the provability of the claim that 'all instances of the axiom schemata of O are true' within CT in Chapter 6. Shapiro in [Shapiro 2002], for example, seem to forget that we do need instances with the truth predicate of the schemata themselves to obtain the claim.

[31] For a detailed proof, see [Halbach 2011, pp. 104–106].

[32] Cfr. [Halbach 2011, Lemma 8.4].

by the following instance of induction of CT:[33]

$$\forall x\big(\forall y < x((\mathsf{P\dot{r}v_{PA}}(y) \to T\mathsf{e\dot{n}d}(y)) \to (\mathsf{P\dot{r}v_{PA}}(x) \to T\mathsf{e\dot{n}d}(x)))\big) \to \forall x(\mathsf{P\dot{r}v_{PA}}(x) \to T\mathsf{e\dot{n}d}(x)). \tag{2.12}$$

In the base case $\mathsf{e\dot{n}d}(x)$ is a basic axiom of PA or an instance of induction of PA, thus it is covered by the considerations above. The induction step, for various choices of the logical calculus, follows from the compositional axioms.

Nonetheless, as noticed by [Heck 2009], something stronger is true: let $\mathsf{CT}[I\Sigma_1]$ be constructed exactly as CT[O], that is with the truth predicate allowed to appear into the Σ_1-induction schema, then

Fact 2.2. $\mathsf{CT}[I\Sigma_1] \vdash \mathsf{GRP_{PA}}$.

The possibility of Fact 2.2 relies on the fact that (2.11) and (2.12) are actually instances of Σ_1-induction. This is immediate by looking at our construction of the syntactic predicates and functions appearing in them. A fortiori, we have that $\mathsf{CT}[I\Sigma_1]$ proves the consistency of each $I\Sigma_i$.

A similar phenomenon arises for different choices of $\mathcal{L}_\mathsf{O}$. Let us consider for instance the case in which $\mathcal{L}_\mathsf{O} = \{\in\}$ and Zermelo Fraenkel set theory is the object theory. As [Fujimoto 2012] shows, the axiomatisation of Tarskian truth over ZF in the style of Table 2.1, that we will call CT[ZF], is formulated in a proper class size language $\mathcal{L}_\in^\infty$ containing constants c_x for all elements of $\mathbb{V}$—standardly defined as $\bigcup_{\xi\in\mathsf{On}} V_\xi$. As in the arithmetical case, we can obtain $\Delta_1^{\mathsf{KP}\omega}$ predicates and functions representing the relevant syntactic relations and operations on $\mathcal{L}_\in^\infty$.[34] In particular, to every constant c_x will correspond a set $\dot{x}$, defined as $(\ulcorner c \urcorner, x)$ where $\ulcorner c \urcorner$ is some fixed finite ordinal.

The axioms of CT[ZF] will take a similar form to the one suggested by Table 2.1.[35] For

[33] Since the induction schema and its course-of-value version are equivalent over PA.

[34] For the details, we refer to [Devlin 1984] and [Mathias 2006].

[35] For a presentation of CT[ZF] (called there TC), see [Fujimoto 2012, §4].

instance, the axioms for atomic formulas of $\mathcal{L}_{\in}^{\infty}$ will be:

$$(\mathsf{CT}\in) \qquad \forall x, y(\mathsf{Sent}_{\mathcal{L}_O^\infty}(\dot{x}\underset{\cdot}{\square}\dot{y}) \to (T\dot{x}\underset{\cdot}{\square}\dot{y} \leftrightarrow x \square y))$$

where $\square$ is in $\{\in, =\}$. We denote with $\mathcal{L}_T$ the language $\mathcal{L}_{\in} \cup \{T\}$. We also introduce some further schemata, considered by Fujimoto already, and dispayed here in Table 2.2.

Sep^+	$\forall x \exists y \forall u\big(u \in y \leftrightarrow u \in x \wedge \varphi(u)\big)$ for all formulas φ of $\mathcal{L}_T$;
$\mathsf{Sep}^{\in}$	$\forall v \forall x\big(\mathsf{S\underset{\cdot}{e}nt}_{\mathcal{L}_{\in}^\infty}(\underset{\cdot}{\forall} v x) \to \forall y \exists z \forall u(u \in z \leftrightarrow u \in y \wedge Tx(\dot{u}/v))\big)$;
Rep^+	$\forall x\Big((\forall u \in x)\exists! v(\varphi(u,v)) \to \exists y \forall v(v \in y \leftrightarrow (\exists u \in x)\varphi(u,v))\Big)$ for all formulas φ of $\mathcal{L}_T$;
$\mathsf{Rep}^{\in}$	$\forall x\Big(\forall w_1 \forall w_2 \forall z\big(\mathsf{S\underset{\cdot}{e}nt}_{\mathcal{L}_{\in}^\infty}(\underset{\cdot}{\forall} w_1(\underset{\cdot}{\forall} w_2 z)) \to$ $(\forall u \in x)\exists! v Tz(\dot{u}/u, \dot{v}/v) \to \exists y \forall v(v \in y \leftrightarrow (\exists u \in x)Tz(\dot{u}/u, \dot{v}/v))\big)\Big)$
Ind^+	$\forall x\big((\forall y \in x)(\varphi(y) \to \varphi(x))\big) \to \forall x \varphi(x)$ for all formulas φ of $\mathcal{L}_T$;
$\omega\mathsf{Ind}^+$	$(\forall x \in \omega)\big((\forall y \in x)(\varphi(y) \to \varphi(x)\big) \to (\forall x \in \omega)\, \varphi(x)$ for all formulas φ of $\mathcal{L}_T$;

Table 2.2: Some schemata of CT[ZF].

As before, we define the Global Reflection Principle for ZF:

$$(\mathsf{GRP}_{\mathsf{ZF}}) \qquad \forall x(\mathsf{S\underset{\cdot}{e}nt}_{\mathcal{L}_{\in}^\infty}(x) \wedge \mathsf{B\underset{\cdot}{e}w}_{\mathsf{ZF}}(x) \to Tx),$$

where $\mathsf{B\underset{\cdot}{e}w}_{\mathsf{ZF}}(x)$ is a $\Sigma_1^{\mathsf{KP}\omega}$ provability predicate satisfying the above-mentioned derivability conditions. We have

Fact 2.3. $\mathsf{CT}[\mathsf{ZF}] \vdash \mathsf{GRP}_{\mathsf{ZF}}$.

Since the strategy is parallel to the arithmetical case, we only highlight some noticeable difference that will be useful for later discussion.

The truth of the finitely axioms of ZF before separation and replacement follows from the provability, in CT[ZF], of the T-biconditionals for $\mathcal{L}_{\in}^{\infty}$.[36] $\mathsf{Sep}^{\in}$ and $\mathsf{Rep}^{\in}$, together with the compositional axioms of CT[ZF], suffice to obtain the truth of all the instances of separation and replacement of ZF. An instance of $\omega\mathsf{Ind}^{+}$ applied to a suitable Δ_1-formulas of $\mathcal{L}_{\in}$ constructed along the lines of (2.12) suffices to capture the argument on the length of the derivation in ZF needed to obtain the claim.

Moreover, we let $\mathsf{CT}[\mathsf{KP}\omega]$ be the theory of truth obtained by adding to $\mathsf{KP}\omega$, formulated in $\mathcal{L}_{\in}^{\infty} \cup \{T\}$,[37] compositional axioms in the style of Table 2.1. Since the schemata of Δ_1-separation and Σ_1-replacement for the language $\mathcal{L}_{\in}^{\infty} \cup \{T\}$ are provable in $\mathsf{CT}[\mathsf{KP}\omega]$, we have

Proposition 2.1. $\mathsf{CT}[\mathsf{KP}\omega] \vdash \mathsf{GRP}_{\mathsf{ZF}}$.

As emphasised by Heck,[38] Fact 2.2 and Proposition 2.1 seem to be on the on hand quite surprising, in the sense that by adding a compositional theory of truth over $I\Sigma_1$ or $\mathsf{KP}\omega$ we originally want, following Tarski's requirement, to formalise in it the usual semantic consistency proof for $I\Sigma_1$ or $\mathsf{KP}\omega$ respectively. What is obtained instead, however, is actually a theory that proves the consistency of theories that already formalise the consistency proof for $I\Sigma_1$ and $\mathsf{KP}\omega$.[39]

On the other hand, these results only highlight once more that the addition of the truth predicate over schematic base theories can determine unpleasant consequences. To this respect, it might be worth mentioning a further example (see [Halbach 2011, p.337]): starting with $I\Delta_0^T$, we have (2.11) as a permissible instance of induction. However, if we only add

[36] More precisely, we can prove the uniform Tarski biconditionals in $\mathsf{CT}[\mathsf{ZF}]\restriction$, which is just CT[ZF] with the schemata of separation and replacement of CT[ZF] restricted to $\mathcal{L}_{\in}^{\infty}$.

[37] Here it suffices that $\mathcal{L}_{\in}^{\infty}$ contains constants for all objects in the intended domain.

[38] Although this observation is ours.

[39] PA proves $\mathsf{Con}_{I\Sigma_1}$ and similarly ZF proves that $\mathbb{L}_{\omega_1^{CK}}$ is a model of $\mathsf{KP}\omega$.

UTB axioms to $I\Delta_0^T$, we obtain all the instances of induction of PA, that is induction for formulas of any complexity.[40] This can create confusions if one wants to define the reflective closure —that is the result of adding all principles that are implicit in the acceptance of the object theory—of systems such as PRA in which some instances of induction are not permissible as only induction on primitive recursive formulas is justified by finitistic means.

At the root of the results sketched above and of Halbach's examples lies the property of the truth predicate of determining a 'collapse' in the arithmetical or Levy's hierachy, so that as soon as a truth predicate is prefixed to them, formulas of any complexity can be inserted into the schemata and treated as terms in atomic formulas. Nonetheless, although technically rather trivial, these facts seem to be quite relevant for our purpose, as we shall argue shortly. Deeper problems are related to the extension of schemata of the object theory to philosophical vocabulary: for further details, we refer to [Halbach 2011].

If the set of axioms of the object theory O is just an arbitrary recursively enumerable set, and not a finite set of formulas of $\mathcal{L}_O$ together with a finite set of schemata in the sense explained above, we cannot in general prove the claim 'all axioms of O are true'. PRA itself, for instance, cannot be finitely axiomatised; in particular, the induction schema for primitive recursive formulas proper of its usual axiomatisations can be eliminated—see [Troelstra&Schwichtenberg 2000, p.128ff]—but the set of axioms characterising the infinitely many symbols for primitive recursive functions in its language cannot be proved

[40] Assuming the antecedent of an arbitrary instance of induction of any complexity

$$\varphi(0) \wedge \forall y(\varphi(y) \rightarrow \varphi(Sy)),$$

we have the following instance of the induction of $I\Delta_0^T$—which is in fact (2.11)—:

$$T\ulcorner\varphi(0)\urcorner \wedge \forall y(T\ulcorner\varphi(\dot{y})\urcorner \rightarrow T(\ulcorner\varphi(\dot{Sy})\urcorner)) \rightarrow \forall y T\ulcorner\varphi(\dot{y})\urcorner$$

By the UTB axioms, this is equivalent to

$$\varphi(0) \wedge \forall y(\varphi(y) \rightarrow \varphi(Sy)) \rightarrow \forall y\varphi(y),$$

then $\forall y\varphi(y)$.

true by merely employing the T-sentences provable in CT[PRA].[41]

In the next section we shall see how the particularities related to the axiomatisation of the truth predicate over schematically axiomatised base theories can support Heck's and Halbach's diagnosis of a confusion between syntactic and mathematical versions of the schemata in the usual construction.

2.3.3 Syntactic and Mathematical Schematic Reasoning

The observations presented above indicate a firm basis to start with: $\mathsf{GRP}_{\mathsf{PA}}$ is only provable in CT because the theory of truth bearers—that is the syntactic, or 'morphological' portion of the Tarskian metatheory—and the mathematical object theory are identified. Only in a setting so-constructed, in fact, it is possible to prove that the schemata of the object theory are true at each substitutional instance. By employing a rather effective slogan attributed to Albert Visser, the provability of $\mathsf{GRP}_{\mathsf{PA}}$ in CT[PA] is a 'happy accident'.

Let us consider the instance (2.11) of the induction schema of CT[PA]: it states that if the formula $\varphi(v)$ is true at the value 0 of its free variable, and that if it is true at the value n then it is true at the value $n+1$, then it is true at each n. Now let us compare it to (2.12), in which it is established some property of an object belonging to our metatheoretic reflection *on* O, that is a proof. There in fact we want express that if an axiom of O and all the (universal closures) of formulas obtained from the axioms by means of the rules of inference of the chosen logical calculus belong to the class of true sentences, then everything that is provable in O also belongs to that class.

In order to express (2.12), in principle, we don't need the induction of PA extended: the induction principle that we need might in fact well belong to a subtheory of PA formalising syntactic and logical notions concerning $\mathcal{L}_{\mathsf{PA}}$, $I\Sigma_1$ for instance, and Proposition 2.2 tells us that an extended instance of Σ_1-induction would suffice to establish the claim. The fact that

[41] For a further example considering the language of analysis, see [Franzel 2004, p.205].

$I\Sigma_1$ and PA are formulated in the same language, as well as the fact that $I\Sigma_1$ is a subtheory of PA, do not have to distract us: they are in principle two different theories designed for radically different purposes in the context of our metamathematical reflection.

The case of CT[ZF] is particularly instructive: it is easy to see that the following holds

$$\mathsf{CT[ZF]}\!\restriction + \mathsf{Sep}^+ + \mathsf{Rep}^+ + \omega\mathsf{Ind}^+ \vdash \mathsf{GRP}_{ZF}. \tag{2.13}$$

(2.13) shows how different are the arguments leading to the provability of the truth of all instances of separation and replacement on the one hand, which requires Sep^+ and Rep^+, and the argument on the length of the derivation needed to perform the inductive step of the proof of Fact 2.3 on the other, for which $\omega\mathsf{Ind}^+$ suffices. The case of CT[ZF] is even more persuasive than the case of CT in this respect: performing the two arguments, in CT, required two instances of the same schema. In the proof of Fact 2.3 we can distinguish between schemata involving the truth predicate employed in a 'mathematical way', such as Sep^+ and Rep^+, and principles belonging to the 'syntactic' component of our metatheoretic picture, such as the extended ω-induction schema needed in the proof.

If one carefully distinguishes between patterns of reasoning belonging to our metamathematical reflection *on* the object theory—in the case at hand ZF—as we shall see in later chapters, it is possible to see how instances of its schemata of the form Sep^+ and Rep^+ and the relevant versions of the schema $\omega\mathsf{Ind}^+$ just considered might even be taken to belong to different theories.

Recap and Things to Come. So far we have tried to reconstruct the route that from one's choice of primary truth bearers, made at the level of philosophical enquiry, leads to the class of objects that formal theories of truth recognise as the objects to which the truth predicate is ascribed. We have also seen how the widespread habit of identifying the theory of those objects with the object theory in the construction of axiomatic theories of truth determines what we called 'third reduction' in Figure 2.1.

We showed that this is not what Tarski had originally in mind: Tarski's Theorem, a later addition to the *Wahrheitsbegriff*, rests upon the possibility of this reduction, whereas the initial picture of the metatheory does not. Moreover, although mathematically rather fruitful, the equation theory of syntax=object theory seems to conceal the subtle distinction between patterns of reasoning belonging to different areas of our metamathematical reasoning and which are clearly distinguished in informal metemathematical practice.[42] This is clear by looking at the axiomatisations of Tarskian truth over schematically axiomatised theories found in the literature.

We will then investigate an axiomatisation of Tarskian truth which appears to be more faithful to the original picture of the metatheory presented in the *Wahrheitsbegriff*. In particular, we will construct our theory of truth distinguishing between the domain of discourse of the syntax theory and a domain of discourse of the object theory. This will be carried out in Chapter 4.

In the next chapter, we are going to introduce a theory that appears to be suitable to characterise the class of objects to which our semantic machinery will be applied.

[42] This point will be extensively treated in Chapter 4.

3 Syntax and Finite Sets

The construction of a truth theory for an object theory O commits us to the existence of objects to which truth is ascribed or to an isomorphic copy of them—at least if truth is considered as a predicate, as we are assuming. In the usual setting, a characterisation of syntactic and logical notions of the object theory O under consideration, fixed and the level of informal mathematics, is replicated within O itself. As a first step towards the main construction investigated in the present work, we propose to fix at outset also the formal system capturing the informal description of the syntax for O, which in turn can remain unspecified, at least at the present stage. As a result, the generality ascribed by Tarski to his method can be partially recovered.

3.1 The Informal Picture

As commonly conceived,[1] primary objects of investigation of the informal theory dealing with the 'syntax' of O are primitive symbols of an arbitrary first-order language $\mathcal{L}$; we stipulate they are $v, c, f, R, \neg, \rightarrow, \forall$. We also assume the capability of talking about natural numbers to index primitive symbols, and the possibility of constructing ordered pairs, triples and tuples of these objects, as highlighted in Table 3.1, so that complex expressions can be formed.

Ordered pairs (a, b) are governed by the usual law and more generally tuples can be

[1]Cfr. for instance [Feferman 1991, Boolos 1993].

$v_i = (v, i)$	$c_x = (c, x)$	$f^i_j = (f, (j, i))$	$R^i_j = (R, (j, i))$
$f^i_j z = (f^i_j, z)$	$R^i_j z = (R^i_j, z)$	where $z = (z_1, ..., z_i)$	
$(\neg a) = (\neg, a)$	$(a \to b) = (\to, (a, b))$	$(\forall x)a = (\forall, (x, a))$	

Table 3.1

defined in terms of pairs as $(a_1, ..., a_n) = (a_1, (a_2, ..., a_n))$. It is worth noticing, although immediate from the definitions , that this representation of complex expressions in terms of tuples gives us unique readability of $\mathcal{L}$-expressions in one step.

Finally we want to talk, in our informal syntax theory for O, about sets of expressions of the language $\mathcal{L}_\mathsf{O}$. Still focusing on an arbitrary $\mathcal{L} \supset \mathcal{L}_\mathsf{O}$, we set:

Var := the set of all v_i where i is a natural number;

Fun := the set of all f^i_j, with i, j are natural numbers and $i \geq 1$;

Rel := the set of all R^i_j, with i, j are natural numbers and $i \geq 1$;

Const := the set of all c_x with x an element in the intended domain of O.

If we want to restrict the attention to the language $\mathcal{L}_\mathsf{O}$ of our object theory and $\mathcal{L}_\mathsf{O} = \mathcal{L}_\mathsf{A} = \{0, S, +, \times\}$ and in our official definition equality is just R^2_1, then $\mathcal{L}_\mathsf{O}$ will just be a subset of $\mathsf{Rel} \cup \mathsf{Fun} \cup \mathsf{Const}$ also including R^2_1. Thus we can relativize the sets below to the language $\mathcal{L}$, so that, for instance, $\mathsf{Fun}_{\mathcal{L}_\mathsf{O}} = \{S, +, \times\} \cap \mathsf{Fun}$ and similarly for other relevant sets.

Our metatheoretic characterisation of the syntax of O also commits us to the notion of finite sequence. A finite sequence s is usually associated with a natural number n, its length, and with its entries s_i, the values of s at each natural number $i < n$. We usually stipulate the existence of the empty sequence and the possibility of always prolonging a given sequence by one element.[2] The basic principle concerning finite sequences—provable in our informal metatheory—prescribes that two finite sequences are identical if and only if they have the same length and agree on all their values.

t is in the set $\mathsf{Term}_{\mathcal{L}_\mathrm{O}}$ of terms of $\mathcal{L}_\mathrm{O}$ if and only if it is the last entry of a finite sequence all whose values are either (i) members of Var; or (ii) members of Const; (iii) or a pair (f^i_j, z) where $z = (z_1, \dots, z_i)$ and each z_k, with $1 \le k \le i$ is in $\mathsf{Term}_{\mathcal{L}_\mathrm{O}}$ already.

Similarly an atomic formula is a pair of the form (R^i_j, z) where z is as before, for some i, j. Assuming that $\mathcal{L}_\mathrm{O}$ features only $\neg, \rightarrow \forall$ among its logical constants, we say that φ is an element of the set $\mathsf{Fml}_{\mathcal{L}_\mathrm{O}}$ of formulas of $\mathcal{L}_\mathrm{O}$ if and only if it is the last entry of a finite sequence whose values are either members of the set $\mathsf{AtFml}_{\mathcal{L}_\mathrm{O}}$ of atomic formulas or an ordered pair of $\neg$ and a previous value of the sequence, or a triple of $\rightarrow$ and two earlier values of the sequence, or a triple of $\forall$, a member of Var and an earlier member of the sequence.

A term t is closed if and only if there is no value in its formation sequence which is an element of Var. An element v of Var is said to be free in the formula φ if and only if there is a finite sequence $\langle s_i : i < k \rangle$ whose last entry is φ, s_0 is of the form $(R^n_j, z_1, \dots, z_n)$ and v occurs in one of the z_m with $m \le n$, and such that for all $j < k$, s_{j+1} is of the form $(\neg, s_j)$, $(\rightarrow, s_j, b)$ or $(\rightarrow, b, s_j)$ for some formula b or s_{j+1} is $(\forall, w, s_j)$ with $w \in \mathsf{Var}$ and $w \neq v$. A sentence thus is just a formula in which no variable is free.

The result of substituting all free occurrences of the variable v for the term r into the term t, denoted with $t(r/v)$, is obtained by taking a construction sequence s for t and generating the formation sequence s' for $t(r/v)$ by examining each value s_i with $i < \mathsf{lh}(s)\big(= \mathsf{lh}(r)\big)$: if s_i is v, then $s'{}_i$ becomes t; otherwise, we identify s_i and s'_i.

The result of substituting all free occurrences of the variable v for the term t in the atomic formula $R^n_j(t_1, \dots, t_n)$ is obtained by replacing each term t_k with $k \le n$ in $R^n_j(t_1, \dots, t_n)$ with $t_k(t/v)$. If ψ is a formula of $\mathcal{L}_\mathrm{O}$ then, $\psi(t/v)$ is defined as the last entry of the sequence s' constructed from the formation sequence s of ψ, say of length m, in the following way: for

[2] In the terminology that will be employed in Chapter 5, we require $\mathsf{Informal\ Morphology}^\mathrm{O}$ to be sequential.

all $k < m$, if s_k is atomic, that is of the form $(R^n_j, t_1, ..., t_n)$ for some n, j, then s'_k is $s_k(t/v)$; if s_k is $(\neg, s_l)$ for some $l < k$, then $s'_k = (\neg, s'_l)$; if s_k is $(\rightarrow, s_l, s_m)$ for some $l, m < k$, then s'_k is $(\rightarrow, s'_l, s'_m)$; finally, if s_k is $(\forall, u, s_l)$ for some u in Var, $u \neq v$ and $l < k$, then s'_k is $(\forall, u, s'_l)$, whereas if s_k is $(\forall, v, s_l)$ for some $l < k$, then s'_k is just s_k.

In order to complete our informal characterisation of the object theory O we should give an account of what a derivation from the axioms of O is.[3] We start with a set $\mathsf{Ax_O}$ compounded by the axioms of the chosen system of (first-order) logic and the specific axioms of O.[4] Given two formulas φ and ψ, we say that ψ is a consequence by Modus Ponens of $(\rightarrow, \varphi, \psi)$ and φ; moreover, $(\forall, v, \varphi)$ is a generalisation of the formula φ. A *proof* in O of a formula φ is thus a finite sequence s such that for any s_i with $i < \mathsf{lh}(s)$, s_i is either an axiom of O or it is obtained from some previous values of the sequence by application of modus ponens and generalisation.

In the next section we are going to introduce an axiomatisation of hereditarily finite sets which naturally captures this informal picture, displaying some advantages over the usual formalisation in arithmetic.

3.2 HF

The system HF, introduced by [Świerczkowski 2003],[5] is formulated in the first-order language $\mathcal{L}_{\mathsf{HF}} = \{0, \triangleleft, \in\}$, where 0 is a constant symbol, $\triangleleft$ is a binary function symbol and $\in$ the standard binary relation symbol for membership. Its axioms are:

The expression $k \triangleleft l$ can be intuitively characterised as saying 'k eats l'. In the language $\mathcal{L}_{\in}$ of set theory, this is equivalent to $k \cup \{l\}$. Given the presence of the constant 0 and of the function symbol $\triangleleft$, the language of HF will include terms. *Closed terms* are thus just

[3] Actually, it would be better to talk about the object *system* O, intended as a set of axioms and rules set up to prove theorems.

[4] For the chosen calculus, see Table 3.4 below.

[5] Similar axiomatisations are also studied in [Montagna&Mancini 1994], [Previale 1994] and [Kirby 2009].

HF1 $\forall k(k = 0 \leftrightarrow \forall l(l \notin k))$
HF2 $\forall k, l, m(k = l \triangleleft m \leftrightarrow \forall n(n \in k \leftrightarrow n \in l \vee n = m))$
HF3 $\varphi(0) \wedge \forall k \forall l(\varphi(k) \wedge \varphi(l) \rightarrow \varphi(k \triangleleft l)) \rightarrow \forall k \varphi(k)$ for any $\varphi(v)$ in $\mathcal{L}_{\mathsf{HF}}$

terms in which variables have been replaced by 0. Terms are formed exclusively by repeated applications of the adjunction operator. For every set k, the *singleton* $\{k\}$ is defined as $0 \triangleleft k$ and thus it is a term. Similarly,

$$\{k, l\} := (0 \triangleleft k) \triangleleft l, \qquad \{k, l, m\} := ((0 \triangleleft k) \triangleleft l) \triangleleft m, \qquad \dots$$

are also terms. The ordered pair $(k, l) := \{\{k\}, \{k, l\}\}$, as well as, more generally, n-tuples, are obviously also terms.

We know sketch the basics of HF to render our presentation basically self-contained. All definitions and proofs are extracted from Świerczkowski's paper, except for some notational variations.

Lemma 3.1 (Extensionality)**.** $\forall u(u \in x \leftrightarrow u \in y) \leftrightarrow x = y$.

Proof. By HF3 to x in the above formula. *qed.*

We assume, without proof, that for any sets x, y there are sets $x \cup y$ and $\bigcup x$.

Lemma 3.2 (Separation)**.** $\forall x \exists y \forall u(u \in y \leftrightarrow u \in x \wedge \varphi(u))$, with y not free in $\varphi(u)$.

Proof. By HF3 applied to x. If $x = 0$, then the required y is 0. For the induction step, the sets $\{u : u \in x \wedge \varphi(u)\}$ and $\{u : u \in z \wedge \varphi(u)\}$ exist. Call the former v. The required y is thus $v \triangleleft z$ if $\varphi(z)$ and v itself if $\neg\varphi(z)$. *qed.*

Lemma 3.3 (Replacement)**.**

$$\forall x\big[(\forall u \in x)\exists! v\ \varphi(u, v) \rightarrow \exists y \forall v(v \in y \leftrightarrow (\exists u \in x)\ \varphi(u, v))\big].$$

Proof. Let $\varphi(u, v)$ be functional. HF3 is now applied to x. For the base case, the required y is 0. Assuming that the range ran_1 of the function defined by $\varphi(u, v)$ with $u \in x$ exists, it has to be proved that the range of the function defined by $\varphi(u, v)$ exists when $u \in x \lhd y$. Call w the unique v such that $\varphi(z, v)$: that is so by assumption. The required set is $\mathsf{ran}_1 \lhd w$. *qed.*

Let $x \subseteq y$ be defined in the usual fashion, and let $x \subset y$ be $x \subseteq y \wedge x \neq y$.

Lemma 3.4 (Power Set)**.** $\forall x \exists y \forall u(u \in y \leftrightarrow u \subseteq x)$.

Proof. The claim is obtained by applying HF3 to x. If $x = 0$, then the required y is $\{0\}$. By assumption, the power set $\mathcal{P}x$ of x exists (cfr. Table 3.2), the set required by the induction step is

$$\mathcal{P}x \cup \{u : (\exists v \in \mathcal{P}x)\ u = v \lhd z\},$$

that is the set obtained from $\mathcal{P}x$ by enlarging it with the result of the adjunction of z to each of its members. *qed.*

The following definitions, although standard, will be useful in what follows. An n-ary *relation* R is a collection of n-tuples. The *domain* of the binary *relation* R is the set

$$\mathsf{dom}(R) := \{u \in \bigcup\bigcup R : \exists v((u, v) \in R)\}, \tag{3.1}$$

whereas the *range* of R is the set

$$\mathsf{ran}(R) := \{v \in \bigcup\bigcup R : \exists u((u, v) \in R)\} \tag{3.2}$$

Table 3.2 summarises some noticeable sets whose existence is provable in HF and the function symbols abbreviating them.

Lemma 3.5 (Regularity)**.** $x \neq 0 \to (\exists u \in x)(x \cap u = 0)$.

$x \cup y :=$	the union of x and y;
$\bigcup z :=$	the union of all elements of elements of x;
$\{u : u \in x \wedge \varphi(x)\} :=$	the class of all $u \in x$ such that $\varphi(u)$ (cfr. Lemma 3.2);
$x \cap y :=$	the intersection of x and y (cfr. Lemma 3.2);
$\bigcap x :=$	the set of all u such that $u \in y$ for some $y \in x$ for nonempty x and 0 otherwise;
$\mathcal{P}x :=$	the set of all subsets of x (cfr. Lemma 3.4);
$x^+ :=$	the successor of x, i.e. $x \triangleleft x$;
$x - y :=$	the difference of x and y, namely $\{u : u \in x \wedge u \notin y\}$.

Table 3.2: Abbreviations for noticeable sets.

Proof. The proof is by induction on the y formula

$$(\forall u \in x)(u \cap x \neq 0) \rightarrow y \notin x \wedge y \cap x = 0. \tag{3.3}$$

qed.

Ordinals are standardly defined as transitive sets all whose elements are transitive—we write $\mathsf{Ord}(x)$ for 'x is an ordinal'. We employ $\alpha, \beta, \gamma, \ldots$ to range over ordinals.[6] In HF, they are obtained from 0 by self-application of the adjunction operator, so that α^+, the successor of α, is defined as $\alpha \triangleleft \alpha$. Occasionally, we will also need to refer to the predecessor of α, which will be denoted by $\overline{\alpha}$. Membership among ordinals, which enjoys the trichotomy

[6] An exception is represented by our coding of variables: if v_i is the i^{th} variable, then its ordinal code will be denoted by $i + 1$, where $+$ is ordinal addition and i a finite ordinal.

property, will be abbreviate as

$$\alpha < \beta :\leftrightarrow \alpha \in \beta;$$
$$\alpha \leq \beta :\leftrightarrow \alpha < \beta \vee \alpha = \beta.$$

Reasoning with finite sets enables one to define, for any set x, a linear order of its elements and the existence of its minimal and maximal elements. This task is easily achieved for sets of ordinals, as the next proposition will show. Later we will consider a generalisation of this order, which holds among hereditarily finite sets in general.

Proposition 3.1.

(i) Every set of ordinals is linearly ordered by <;

(ii) Every nonempty set of ordinals has a maximal and a minimal element.

Proof. (i) is immediate.

The existence of the minimal element $\mathsf{min}_<(x)$ (and symmetrically for $\mathsf{max}_<(x)$) of the nonempty set of ordinals x is proved by applying HF3 to the x in

$$x \neq 0 \wedge (\forall \alpha \in x)\mathsf{Ord}(\alpha) \to (\exists \beta \in x)(\forall \alpha \in x)\beta \leq \alpha. \tag{3.4}$$

For the induction step, $\mathsf{min}_<(x \lhd y)$ is either $\mathsf{min}_<(x)$, which exists by induction hypothesis, or it is y itself. *qed.*

Proposition 3.2 (Strong Induction on Ordinals)**.**

$$\forall \alpha((\forall \beta < \alpha)\ (A(\beta) \to A(\alpha))) \to \forall \alpha\ A(\alpha).$$

Proof. By assumption we have $A(0)$. To establish $A(x \lhd y)$ for arbitrary ordinals x, y, we notice that if $x \lhd y$ is an ordinal, then $x = y$. If $\neg A(x \lhd y)$, then the set

$$\{u : \mathsf{Ord}(u) \wedge \neg A(u)\}$$

is nonempty and has a least element α by regularity. But α cannot be 0, nor a successor ordinal by assumption. We can then apply HF3. *qed.*

The standard model of HF. We reason (in a portion of) ZF. For the sake of readability, in the current exposition we distinguish metatheoretic symbols from their counterparts in $\mathcal{L}_{\mathsf{HF}}$ by placing a dot over them.

Definition 3.1 (Constant Terms)**.** The class $\mathbb{C}$ of closed terms is the smallest class containing 0 and closed under $\lhd$, that is for every $x, y \in \mathbb{C}$, also $x \lhd y \in \mathbb{C}$.

Lower case latin letters $r, s, t, \ldots$ will be employed to range over elements of $\mathbb{C}$. We denote with $\equiv$ the equivalence relation on $\mathbb{C}$ defined as

$$s \equiv t :\leftrightarrow \mathsf{HF} \vdash s = t. \tag{3.5}$$

Moreover, we standardly define the equivalence class $[t] = \{s \dot{\in} \mathbb{C} : s \equiv t\}$ and the class $\mathbb{H}$, which will be the domain of our model $\mathcal{H}$,

$$\mathbb{H} = \{[t] : t \dot{\in} \mathbb{C}\}, \tag{3.6}$$

that is the set of all equivalence classes of closed terms. The interpretation of $\in$ is defined as

$$\text{for all } s, t \in \mathbb{C}, \qquad [s] \dot{\in} [t] :\leftrightarrow \mathsf{HF} \vdash s \in t, \tag{3.7}$$

and the interpretation if $\lhd$ in such a way that for all $t_0, t_1 \in \mathbb{C}$,

$$[t_0] \dot{\lhd} [t_1] = [t_0 \lhd t_1] \tag{3.8}$$

Finally, $[0]$, is abbreviated as 0.

Definition 3.2 (Standard Model)**.** Let $0, \lhd, \in$ interpreted as above. The structure $\mathcal{H} = (\mathbb{H}, 0, \dot{\lhd}, \dot{\in})$ is the standard model of HF.

For the (straightforward) proof that $\mathcal{H}$ satisfies the axioms of HF, provided that the latter is consistent, see [Świerczkowski 2003, pp. 56-57].

Σ_1-formulas. Due to the presence of terms in the language of HF, we need to give a slightly modified definition of the Levy's hierarchy. For our purposes, it suffices to introduce the notion of *strict* Σ_1-formula.

Definition 3.3 (Strict Σ_1-formula)**.** The class of strict Σ_1-formulas is the smallest class of formulas of $\mathcal{L}_{\mathsf{HF}}$ including atomic formulas $v_i \in v_j$ and $v_i = v_j$ where v_i and v_j are variables, and closed under $\vee$, $\wedge$, bounded universal quantification $(\forall v_i \in v_j)\varphi$, with φ already strict Σ_1, and $i \neq j$, and existential quantification.

A Σ-*formula* is just a formula of $\mathcal{L}_{\mathsf{HF}}$ which is provably equivalent to a strict Σ_1-formula. Consequently, a Σ-*sentence* is a Σ-formula that is also a sentence. The more general definition of $\Sigma_1^{\lhd}$-formulas of $\mathcal{L}_{\mathsf{HF}}$ thus is built up on the previous one:

Definition 3.4 ($\Sigma_1^{\lhd}$-formulas)**.** The class of $\Sigma_1^{\lhd}$ formulas of $\mathcal{L}_{\mathsf{HF}}$ is defined as the smallest class of formulas containing atomic formulas $s \in t$ and $s = t$ for (not necessarily closed) terms s, t and closed under $\wedge$, $\vee$, bounded universal quantification of the form $\forall v_i(v_i \in t \rightarrow \ldots)$, where t is any term not containing v_i, and unbounded existential quantification.

In the next section we show how to obtain, within HF, predicates corresponding to the informal characterisation of syntactic and logical notions of $\mathcal{L}_{\mathsf{O}}$ given above. The equivalences shown below indicate that the formulas displayed are all Σ_1-formulas.

(3.9) $l \in \mathsf{dom}(k) \leftrightarrow (\exists n \in k)(\exists m \in n)(\exists p \in m)\ n = (l, p);$

(3.10)

$$l \in \mathsf{ran}(k) \leftrightarrow (\exists n \in k)(\exists m \in n)(\exists p \in m)\ n = (p, l);$$

(3.11) $\mathsf{Ord}(k) \leftrightarrow (\forall l \in k)\big[(l \subseteq k) \wedge (\forall m \in l)(m \subseteq l)\big];$

(3.12)

$$\mathsf{Seq}(s, \alpha) \leftrightarrow 0 \in \alpha \wedge (\forall \beta < \alpha)(\exists l \in \mathsf{ran}(s))((\beta, l) \in s) \wedge$$
$$(\forall m, n \in s)(\exists \gamma, \delta \in \alpha)(\exists p, q \in \mathsf{ran}(s))(p = (\gamma, m) \wedge q = (\delta, n) \wedge (\gamma \neq \delta \vee m = n))$$

(3.13) $\mathsf{SeqEnd}(s, \alpha, k) \leftrightarrow \mathsf{Seq}(s, \alpha) \wedge (\exists \beta \in \alpha)(\alpha = \beta \lhd \beta \wedge (\beta, k) \in s).$

$\mathsf{Seq}(s, \alpha)$ thus expresses that s is a sequence of length α; $\mathsf{SeqEnd}(s, \alpha, k)$ that k is the last entry of the sequence s of ordinal length α. Given an $\mathcal{L}_{\mathsf{HF}}$-formula φ, we abbreviate with $\varphi(s_\alpha)$ the formula $\exists k((\alpha, k) \in s \wedge \varphi(k))$.

The notion of Σ_1-formula can be used to define a subsystem of HF obtained by restricting HF3 and studied by Kirby in [Kirby 2009] as a counterpart of the well-known arithmetical system $I\Sigma_1$.

Definition 3.5 (Σ_1HF)**.** Σ_1HF is the subsystem of HF obtained from HF by restricting HF3 to Σ_1-formulas.

The interesting property of Σ_1HF, from a foundational point of view, is that we can prove an analog of Parsons' Theorem in the set theoretic context: primitive recursive set functions on hereditarily finite sets are exactly the functions that can be proved total in Σ_1HF. According to a popular reading, therefore (see [Tait 1981], for instance), Σ_1HF captures finitistic reasoning on sets.

Σ_1-formulas play an important role in metamathematics. The current setting is no exception:

Proposition 3.3 (Σ_1-completeness)**.** Let σ be a Σ_1-sentence, if $\mathcal{H} \vDash \sigma$, then $\mathsf{HF} \vdash \sigma$.

Proof. The result is obtained for $\Sigma_1^{\triangleleft}$-sentences, by noticing that every Σ_1-sentence is also a $\Sigma_1^{\triangleleft}$-sentence. We define the length $\lambda(\varphi)$ of a $\Sigma_1^{\triangleleft}$-formula as the number of occurrences of $\vee, \wedge, \exists$ and $\forall$ in φ. The proof is by metatheoretic induction on the length of the Σ_1-sentence σ.

If $\lambda(\sigma) = 0$, then σ is of the form $s = t$ or $s \in t$ with s, t in the class $\mathbb{C}$ of closed terms. This case follows immediately from (3.5) and (3.7) respectively.

For the induction step, let us assume that the claim holds for all sentences τ such that $\lambda(\tau) < \lambda(\sigma)$. The claim is straightforwardly obtained for the case in which σ is $\tau \wedge \chi$ or $\tau \vee \chi$. If σ is $\exists v_i \varphi$, then by assumption there is a $t \in \mathbb{C}$ such that $\mathcal{H} \vDash \varphi(t/v_i)$. By induction hypothesis, $\mathsf{HF} \vdash \varphi(t)$, and thus $\mathsf{HF} \vdash \exists v_i \varphi$.

Finally, if σ is $(\forall v_i \in t)\varphi$, we first notice that in HF is provable that any closed term t can be decomposed in a finite list of terms $t_0, ..., t_n$, that is

$$(\forall x \in t)\ \varphi(x) \leftrightarrow \varphi(t_1) \wedge ... \wedge \varphi(t_n).$$

We then assume $\mathcal{H} \vDash \varphi(t_1)$,..., $\mathcal{H} \vDash \varphi(t_n)$; by induction hypothesis, $\mathsf{HF} \vdash \varphi(t_1)$,...,$\mathsf{HF} \vdash \varphi(t_n)$, thus $\mathsf{HF} \vdash \varphi(t_0) \wedge ... \wedge \varphi(t_n) \leftrightarrow \sigma$. *qed.*

Definining functions in HF. We introduce two methods for defining new functions from old in HF.

Proposition 3.4 (Recursion on Ordinals)**.** For every closed term t and functions h, g there is a function f such that:

$$\begin{aligned} &f(0) = t; \\ &f(k) = g(f(\overline{k})), \quad \text{if } k \text{ is a nonzero ordinal;} \\ &f(k) = h(k), \quad \text{if } k \text{ is not an ordinal.} \end{aligned} \tag{3.14}$$

Proof. Consider the $\mathcal{L}_{\mathsf{HF}}$-formula

$$\begin{aligned}\varphi(k,l) :\leftrightarrow\ & \textit{either } k \text{ is a nonzero ordinal and} \\ & \text{there is a sequence } s \text{ of length } k \text{ such that: } l = g(s_{\overline{k}}) \wedge \\ & (\forall \alpha < k)(\alpha = 0 \wedge s_\alpha = t) \vee (\alpha \neq 0 \wedge s_\alpha = g(s_{\overline{\alpha}})), \\ & \textit{or } k \text{ is not an ordinal and } l = h(k), \\ & \textit{or } k = 0 \text{ and } l = t.\end{aligned}$$

There is a unique l that satisifies $\varphi(k,l)$. Moreover, $\varphi(k,l)$ satisfies the clauses in (3.14).

qed.

Example. The function $w(k)$ will turn out to be crucial for the proof of Gödel's First Incompleteness Theorem presented in the Appendix to the present chapter. $w(k)$ takes the code of a variable and returns the code of the nonzero ordinal associated with it. It is defined by recursion on ordinals as

$$w(k) = \begin{cases} 0 \quad \text{if } k = 0 \text{ or } k \text{ is not an ordinal;} \\ (\ulcorner \lhd \urcorner, w(\overline{k}), w(\overline{k})) \quad \text{if } k \text{ is an ordinal and } k \neq 0. \end{cases}$$

▪

A second method for generating new functions from old is by employing the partial order of hereditarily finite sets given by their rank. This method differs from the one just considered in the fact that the function f resulting from the recursion is defined on all finite sets and not merely on ordinals. By Proposition 3.4 we define the function ρ as:

$$\rho(0) = 0;$$
$$\rho(k) = \mathcal{P}\rho(\overline{k}) \quad \text{if } k \text{ is a nonzero ordinal;}$$
$$\rho(k) = 0 \quad \text{if } k \text{ is not an ordinal.}$$

Crucially, HF proves that for any k there is an ordinal α such that $k \in \rho(\alpha)$, that is that the function $\rho(.)$ exhausts the universe of finite sets.[7] The rank of a finite set k, denoted by $\mathsf{rank}(k)$, is defined as the least α such that $k \subseteq \rho(\alpha)$. Finally, the transitive closure of k, $\mathsf{TC}(k)$, is standardly defined as the minimal transitive set l such that $k \subseteq l$. The following definition, however, immediately entails the existence of $\mathsf{TC}(k)$ for each k (cfr. Lemma 3.2):

(3.15) $$\mathsf{TC}(k) = \bigcap\{l \in \rho(\alpha) : k \subseteq l \text{ and } l \text{ is transitive}\} \quad \text{for some } \alpha.$$

Let us define the restriction of a function f to l as $f \upharpoonright l = \{(m, f(m)) : m \in l)\}$. The expression of 'definition by recursion on rank' is warranted by the fact that $k \in l$ entails $\mathsf{rank}(k) < \mathsf{rank}(l)$, and $\mathsf{rank}(k) = \mathsf{rank}(\mathsf{TC}(k))$ for all k, then if $k \in \mathsf{TC}(l)$, then also $\mathsf{rank}(k) < \mathsf{rank}(l)$.

Proposition 3.5 (Recursion on rank)**.** For every function $g(k, l)$ on finite sets, there is a function f such that

$$\mathsf{HF} \vdash f(k) = g(k, f \upharpoonright \mathsf{TC}\, k).$$

Proof Idea. Crucially, for any k we can find a function f_k with domain $(\mathsf{TC}\ k) \lhd k$ (see [Świerczkowski 2003, p. 50]). Then one checks that $f_k(l) = f_l(l)$ for all $l \in \mathsf{TC}\ k$.

By adapting the same argument, we obtain:

[7] Cfr. [Świerczkowski 2003], Theorem 4.3.

Corollary 3.1. For every function function $g(k_1, ..., k_n)$ there is a $n-1$ary function $f(k_0, ..., k_{n-1})$ such that

$$\mathsf{HF} \vdash f(k_1, ..., k_{n-1}) = g(k_1, ..., k_{n-1}, f \upharpoonright (\mathsf{TC}(k_1) \times ... \times \mathsf{TC}(k_{n-1}))).$$

Example. We provide an example of a function generated by recursion on rank. Again this function plays a crucial role in the beautiful proof of the Incompleteness Theorems in HF. Let

$$\tau(k, l) := (\ulcorner \triangleleft \urcorner, (\ulcorner \triangleleft \urcorner, 0, (\ulcorner \triangleleft \urcorner, 0, k)), (\ulcorner \triangleleft \urcorner, (\ulcorner \triangleleft \urcorner, 0, k), l)).$$

Later we shall see how coding, also for the language of HF, can be developed in HF. We just anticipate that is k and l code s and t respectively, $\tau(k, l)$ is equal to $\ulcorner(s, t)\urcorner$. Therefore if $k = \ulcorner s \urcorner$ and $l = \ulcorner t \urcorner$ for any closed terms s, t, we that $\tau(k, l) = \ulcorner(s, t)\urcorner$. This enables one to obtain, given the code $\ulcorner \varphi \urcorner$ of a formula, its code $\ulcorner\ulcorner \varphi \urcorner\urcorner$. Then we define by recursion on rank a function

$$h(k) = \begin{cases} w(k) & \text{if } k \text{ is an ordinal;} \\ \tau(h(m), h(n)) & \text{if } k = (m, n); \\ 0 & \text{otherwise.} \end{cases}$$

We notice that for any $k = (m, n)$, clearly m and n are in $\mathsf{TC}\ k$, so we can apply Corollary 3.1. ▪

The order of hereditarily finite sets. Hereditarily finite sets can be partially ordered by the subset relation and by the <-relation applied to ranks. It is provable in HF that hereditarily finite sets enjoy a global linear order which generalises the relations above.

New relations $\varphi(k_1, ..., k_n)$ can be introduced by recursion on rank by considering their characteristic function χ_φ such that

$$\mathsf{HF} \vdash \chi_\varphi(k_1, \ldots, k_n) = 1 \leftrightarrow \varphi(k_1, \ldots, k_n).$$

Now the (global) order relation required is ensured by the following:

Proposition 3.6. There is a relation $<$ such that

$$\mathsf{HF} \vdash k < l \leftrightarrow (\exists m \in l - k)(\forall n \in k - l)(n < m).$$

We refer to [Świerczkowski 2003] for a proof that $<$ is indeed a global order for hereditarily finite sets. We just notice that since any order of a finite set is also a well-ordering, then Proposition 3.6 ensures, similarly to ordinals, that every nonempty set has a $<$-maximal and an $<$-minimal element and that the following induction scheme holds:

$$\forall k((\forall l < k)\varphi(l) \to \varphi(k)) \to \forall k \varphi(k). \tag{3.16}$$

(3.16) also yields another verification of Proposition 3.16.

Let us now define the interval

$$[0, x) := \text{ the unique } z \text{ such that } \forall u(u \in z \leftrightarrow u < x). \tag{3.17}$$

The existence of such a z is guaranteed by Separation. We can thus introduce a further method to define functions in HF: the value of the function $f(x)$ will depend only on a binary function $g(x, y)$ and on the values of $f(y)$ for $y \in [0, x)$. The strategy behind the proof of the following Proposition is similar to the one sketched for Proposition (3.5).

Proposition 3.7 (Definition by Recursion on Inequality)**.** For every function $g(x, y)$ there is a function f such that

$$\mathsf{HF} \vdash f(x) = g(x, f \upharpoonright [0, x)) \tag{3.18}$$

Example. It is easy to define function that, given two sets x, y, yields $(\ulcorner \lhd \urcorner, x, y)$. Now consider the function $i(x)$ defined as follows, where $v = \mathsf{max}_<(x)$ and $u = x - \{v\}$.

$$i(x) = \begin{cases} 0, & \text{if } x = 0; \\ (\ulcorner \lhd \urcorner, i(u), i(v)), \end{cases} \tag{3.19}$$

For the definition to apply, we need to ensure that $u, v < x$. But on the one hand

$$v \in x \to \mathsf{rank}(v) < \mathsf{rank}(x) \to v < x;$$

on the other, we also directly verify by Proposition 3.6 that $u < x$. ▪

3.2.1 The Ackermann Interpretation.

Wilhelm Ackermann in [Ackermann 1937] noticed that if the membership relation is interpreted in a suitable way we can induce a bijection between $\mathbb{H}$ and the standard model $\mathbb{N}$ of arithmetic. This interpretation of the membership relation, known as the *Ackermann Interpretation*, belongs to the logical folklore and it has been thoroughly studied by Kaye and Wong in [Kaye&Wong 2003]. In particular, there it is shown that we can find also an interpretation of Peano Arithmetic into ZF minus the axiom of infinity so that this interpretation and the Ackermann Interpretation are inverse to each other. They are thus synonymous in the sense of [Visser 2008c].

To establish the synonymity of HF and PA, it suffices to verify that the theories definitionally extend each other. In what follows, we will still call Ackermann Interpretation the relative interpretation of HF in PA just mentioned, whereas the interpretation of PA in HF that maps natural numbers into finite ordinals will be called *Ordinal Interpretation*. The inverse of the Ackermann interpretation, which will be presented below, differs form the Ordinal Interpretation and it will be called *Inverse Interpretation*.

The Ackermann interpretation, syntactically, maps the membership relation of $\mathcal{L}_{\mathsf{HF}}$ into the 'formula' of the language of arithmetic:

$$x \text{ occurs as an exponent in the binary representation of } y.$$

More specifically, already in $I\Delta_0$ we can prove the existence and uniqueness conditions of the (partial) function $2^x = y$ (see [Hájek&Pudlák 1993, §I.1.(a) and Ch. V]). Since PA is the interpreting theory, there is no need to worry about the totality of $2^x = y$. We also employ the following:

Fact 3.1. For each x, y there are unique $u \leq y, v \leq 1$ and $w < 2^x$ such that

$$y = 2^{x+1} \cdot u + 2^x \cdot v + w.$$

The proof of the Lemma can be found in [Hájek&Pudlák 1993, Ch. I(b)], where it also shown that the proof can be carried out in $I\Delta_0(\mathsf{exp})$. Let us denote with $\mathsf{bit}(x, y)$ the unique $v \leq 1$ such that

$$(\exists u \leq y)(\exists w < 2^x)(y = 2^{x+1} \cdot u + 2^x \cdot v + w).$$

Let us also recall that a direct interpretation is a relative interpretation between languages in which the identity is preserved and there is no relativisation of quantifiers.[8] The Ackermann interpretation $\mathfrak{a}\colon \mathcal{L}_{\mathsf{HF}} \to \mathcal{L}_{\mathsf{A}}$—where $\mathcal{L}_{\mathsf{A}}$ denotes the language of arithmetic—is a direct

[8]For the details concerning relative interpretations, we refer to §4.4 below.

interpretation such that, crucially:

$$(x \in y)^{\mathfrak{a}} :\leftrightarrow \mathsf{bit}(x, y) = 1. \tag{3.20}$$

Alternatively, we may say that the Ackermann interpretation determines a definitional extension PA⁺ of PA which proves the axioms of HF, obtained by extending PA with the axioms:

PA^+1	$x \in^* y \leftrightarrow \mathsf{bit}(x, y) = 1;$
PA^+2	$x \in^* y \lhd^* z \leftrightarrow x \in^* y \lor x = z.$

defining the new symbols $\in^*$ and $\lhd^*$ which are then provably coextensive with $\in$ and $\lhd$ of $\mathcal{L}_{\mathsf{HF}}$.

Conversely, the Inverse Interpretation $\mathfrak{o}\colon \mathcal{L}_{\mathsf{A}} \to \mathcal{L}_{\mathsf{HF}}$ is defined in the following way. For simplicity, we can work in HF and define there by recursion on the rank of hereditarily finite sets (Cfr. Corollary 3.1) the function $\mathfrak{e}(.)$ such that

$$\mathfrak{e}(k) = \sum_{l \in k} 2^{\mathfrak{e}(l)}, \tag{3.21}$$

which 'enumerates' all finite sets. This function determines a definable bijection, within HF, between finite sets and ordinals. This enables one to definitionally extend HF with definitions of the operations of successor, addition and multiplication so that the resulting theory HF⁺ proves the axioms of PA. In particular those axioms assign to arithmetical successor S the formula $\mathfrak{e}(x) \lhd \mathfrak{e}(x) = \mathfrak{e}(y)$, to the arithmetical addition $+$ the formula $\mathfrak{e}(x) + \mathfrak{e}(y) = \mathfrak{e}(z)$ and similarly for $\times$. To the 'arithmetical' 0 is assigned the constant 0.

Therefore we have the relations shown in Figure 3.1, where the arrows denote inclusion.

Figure 3.1 is of particular importance for some of the results presented in the next chapter.

We can thus summarise the considerations above it in a more perspicuous way:

$$\begin{tikzcd}
 & \mathsf{PA}^+ \arrow[dd, leftrightarrow] & \\
\mathsf{HF} \arrow[ur] \arrow[dr] & & \mathsf{PA} \arrow[ul] \arrow[dl] \\
 & \mathsf{HF}^+ &
\end{tikzcd}$$

Figure 3.1

Fact 3.2. PA and HF are synonymous.

3.3 Capturing the informal picture

As anticipated, we are mainly interested in employing the expressive resources of HF to formalise syntactic and logical notions concerning an arbitrary first order language $\mathcal{L}_{\mathsf{O}}$, as informally depicted in §3.1. We elaborate on Świerczkowski's definitions—which in turn are inspired by Boolos's definitions in [Boolos 1993]—and adapt them to our setting in which $\mathcal{L}_{\mathsf{O}}$ may not be identical to $\mathcal{L}_{\mathsf{HF}}$. Some definitions are new. We elaborate on the reasons for preferring HF to other possible choices in the final section.

Coding. As pointed out in [Świerczkowski 2003], a danger that should be avoided, when coding expressions by finite sets, is to assign two different terms of $\mathcal{L}_{\mathsf{HF}}$ denoting the same set, such as for instance $\{\{0\}, 0\}$ and $\{0, \{0\}\}$, to two distinct expressions of $\mathcal{L}_{\mathsf{O}}$. In other words, if s, t are different closed terms (as strings of symbols) of HF, it is not ensured that HF proves $s \neq t$. However, it is useful in HF to require that, when we talk about the code $\ulcorner \varphi \urcorner$ of the formula φ, for instance, there is a unique closed term in the language naming this code.

The possibility of selecting codes from the class $\mathbb{C}$ of closed terms, therefore, is ruled out.

The attention has to be focused instead on a subset of $\mathbb{C}$.

Definition 3.6 (The set $\mathbb{G}$)**.** Let $\mathbb{G}$ be the smallest class of closed terms such that:

(i) 0 is in $\mathbb{G}$;

(ii) if the closed term s is in $\mathbb{G}$, then $s \lhd s$ is in $\mathbb{G}$.

(iii) If s, t are in $\mathbb{G}$, then (s, t) is in $\mathbb{G}$.

Conditions (i) and (ii) only state that $\mathbb{G}$ contains ordinals. Let us call On this set. Condition (iii) tells us that $\mathbb{G}$ includes also pairs of elments of $\mathbb{G}$. The following proposition tells us that the ambiguity detected among constant terms cannot be replicated if we concentrate on elements of $\mathbb{G}$.

Proposition 3.8. For any terms $s, t \in \mathbb{G}$, if s and t are different expressions of $\mathcal{L}_{\mathsf{HF}}$, then $\mathsf{HF} \vdash s \neq t$.

Proof Sketch. We first notice that for any $s, t \in \mathbb{C}$, either $\mathsf{HF} \vdash s = t$ or $\mathsf{HF} \vdash s \neq t$.[9] Then one checks that it cannot be the case that for distinct $s, t \in \mathbb{C}$, that is with either $\lambda(s) < \lambda(t)$ or $\lambda(t) < \lambda(s)$, we have $\mathsf{HF} \vdash s = t$. If s, t are ordinals or only one of them is an ordinal and the other an ordered pair, this is straightforward. If they are both pairs (s_0, t_0) and (s_1, t_1), we employ the previous cases and the inductive hypothesis to show that s_0 and s_1 and t_0 and t_1 are provably distinct in HF. The result then follows from the law of ordered pairs. *qed.*

To mimic, within HF, the work carried out in §3.1, we start by assigning elements of $\mathbb{G} - \mathsf{On}$ to primitive symbols $v, c, f, R, \neg, \to, \forall$ of $\mathcal{L}_{\mathsf{O}}$. It is not necessary to specify particular sets: we refer to them as $\underset{\cdot}{v}, \underset{\cdot}{c}, \underset{\cdot}{f}, \underset{\cdot}{R}, \underset{\cdot}{\neg}, \underset{\cdot}{\to}, \underset{\cdot}{\forall}$ respectively. With the help of ordinals, we formalise the construction of indexed symbols; the variable v_α, with $\alpha > 0$ will be represented by the ordinal α.

$$\ulcorner c_\alpha \urcorner = (\underset{\cdot}{c}, \alpha), \qquad \ulcorner f_\alpha^\beta \urcorner = (\underset{\cdot}{f}, (\alpha, \beta)), \qquad \ulcorner R_\alpha^\beta \urcorner = (\underset{\cdot}{R}, (\alpha, \beta)).$$

[9] Cfr. [Świerczkowski 2003, Corollary 6.4].

Notational Remark. In what follows we will sometimes employ $1, 2, ..i., n$ to index primitive symbols, variables in particular: however, it should be kept in mind that those variables range over ordinals in the sense of HF.

The semiformal characterisation of complex expressions in terms of tuples given in the opening section can now be easily formalised in HF.

For every terms $t_1, \ldots, t_n$ of $\mathcal{L}_{\mathsf{O}}$	$\ulcorner f^n_i(t_1, \ldots, t_n)\urcorner = (\ulcorner f^n_i\urcorner, \ulcorner t_1\urcorner, \ldots, \ulcorner t_n\urcorner) = \underset{\cdot}{f}{}^n_i t_1, \ldots, t_n$;
	$\ulcorner R^n_i(t_1, \ldots, t_n)\urcorner = (\ulcorner R^n_i\urcorner, \ulcorner t_1\urcorner, \ldots, \ulcorner t_n\urcorner) = \underset{\cdot}{R}{}^n_i t_1, \ldots, t_n$.
For every formulas φ, ψ of $\mathcal{L}_{\mathsf{O}}$,	$\ulcorner \neg\varphi\urcorner = (\ulcorner\neg\urcorner, \ulcorner\varphi\urcorner) = \underset{\cdot}{\neg}\ulcorner\varphi\urcorner$;
	$\ulcorner \varphi \to \psi\urcorner = (\ulcorner\to\urcorner, \ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) = \ulcorner\varphi\urcorner \underset{\cdot}{\to} \ulcorner\psi\urcorner$;
	$\ulcorner(\forall v_i)\varphi\urcorner = (\ulcorner\forall\urcorner, \ulcorner v_i\urcorner, \ulcorner\varphi\urcorner) = \underset{\cdot}{\forall} v_i \ulcorner\varphi\urcorner$.

Table 3.3: Codes of primitive expressions of $\mathcal{L}_{\mathsf{O}}$.

The definitions enable us to introduce the following extra-notation, which will be useful later in the chapter:

$$\mathsf{Neg}(k) := \underset{\cdot}{\neg} k \qquad\qquad \mathsf{Imp}(k, l) := k \underset{\cdot}{\to} l$$

$$\mathsf{All}(i, k) := \underset{\cdot}{\forall} v_i k \qquad\qquad \mathsf{Eq}(k, l) := \underset{\cdot}{R}{}^2_0 k, l, \text{ that is } \ulcorner k \doteq l\urcorner.$$

In the languages considered below, we can always assume that R^2_0 is denoted by $\doteq$ and thus that the definition of $\mathsf{Eq}(k, l)$ obtains.

Codes of variables are just nonzero ordinals, and constants are pairs whose first entry is

$(0, 0)$:

$$\mathsf{var}_{\mathcal{L}_O}(k) :\leftrightarrow 0 < k \wedge \mathsf{Ord}(k);$$
$$\mathsf{const}_{\mathcal{L}_O}(k) :\leftrightarrow \exists l(k = (\c{c}, l)).$$

We recall that a term of $\mathcal{L}_O$ is defined as the last entry of a formation sequence whose elements are either codes of variables, or codes of constants, or tuples whose first element is a code for a function symbol of $\mathcal{L}_O$ and the remaining components are earlier values of the formation sequence. So we first define

$$\mathsf{TermSeq}_{\mathcal{L}_O}(s, \alpha, k) :\leftrightarrow \mathsf{SeqEnd}(s, \alpha, k) \wedge (\forall \beta \in \alpha)\Big(\mathsf{var}_{\mathcal{L}_O}(s_\beta) \vee \mathsf{const}(s_\beta) \vee$$
$$(\exists \gamma, \delta)(\exists \beta_1, \ldots, \beta_\delta \in \beta)(s_\beta = (\ulcorner f^\delta_\gamma \urcorner, s_{\beta_1}, \ldots, s_{\beta_\delta}))\Big).$$

In what follows, instead of employing $\mathsf{TermSeq}_{\mathcal{L}_O}(s, \alpha, k)$, we will instead make us of the Σ_1-formula $\mathsf{TermSeq}^*_{\mathcal{L}_O}(s, \alpha, k)$; it is a slight modification of the previous one and states that every element of the formation sequence for the term coded by k is a subterm of it.[10] We let

$$\mathsf{Term}_{\mathcal{L}_O}(k) :\leftrightarrow \exists s \exists \alpha \mathsf{TermSeq}^*_{\mathcal{L}_O}(s, \alpha, k),$$

which is indeed a strict Σ_1-formula if $\mathsf{TermSeq}^*_{\mathcal{L}_O}(s, \alpha, k)$ is.

To define the notion of being an formula of $\mathcal{L}_O$, we first consider atomic formulas (let us

---

[10] $\mathsf{TermSeq}^*_{\mathcal{L}_O}(s, \alpha, k)$ is obtained simply by adding to $\mathsf{TermSeq}_{\mathcal{L}_O}(s, \alpha, k)$ a clause stating that every s_β, with $\beta \in \alpha$, either is the least element of the sequence, that is $\beta = \overline{\alpha}$, or it is employed in the construction of some s_m with $m > l$. In other words, s_l if it is not the last entry of s has to code some term that appears in some tuple coding a function expression of $\mathcal{L}_O$. The definition would be incredibly long, although stragithforward, and we omit it.

recall that t_i ranges closed terms of HF, whereas $\mathfrak{t}_i$ ranges over all terms of $\mathcal{L}_{\mathsf{O}}$):

$$\mathsf{AtFml}_{\mathcal{L}_{\mathsf{O}}}(k) :\leftrightarrow (\exists \beta, \gamma)(\exists k_1, \ldots, k_\gamma)(\mathsf{Term}_{\mathcal{L}_{\mathsf{O}}}(k_1) \wedge \ldots \wedge \mathsf{Term}_{\mathcal{L}_{\mathsf{O}}}(k_\gamma) \wedge \\ k = (\ulcorner R^{\gamma}_{\beta} \urcorner, k_1, \ldots, k_\gamma)).$$

Now we come to the notion of being a formation sequence of a formula of $\mathcal{L}_{\mathsf{O}}$. We first define the $\mathcal{L}_{\mathsf{HF}}$-formula expressing the construction of an $\mathcal{L}_{\mathsf{O}}$-formula (coded by) m from formulas n, p and a variable i by a single application of $\neg, \rightarrow, \forall$:

$$\mathsf{AppFml}_{\mathcal{L}_{\mathsf{O}}}(i, n, p, m) :\leftrightarrow \mathsf{Var}(i) \wedge \\ \big(m = \neg n \vee m = n \underset{\cdot}{\rightarrow} p \vee m = p \underset{\cdot}{\rightarrow} n \vee m = \mathsf{All}(i, n)\big).$$

Therefore we have:

$$\mathsf{FormSeq}_{\mathcal{L}_{\mathsf{O}}}(s, \alpha, k) :\leftrightarrow \mathsf{SeqEnd}(s, \alpha, k) \wedge (\forall \beta \in \alpha)(\mathsf{AtFml}_{\mathcal{L}_{\mathsf{O}}}(s_\beta) \vee \\ (\exists \gamma, \delta \in \beta)(\exists i)(\mathsf{AppFml}_{\mathcal{L}_{\mathsf{O}}}(i, s_\gamma, s_\delta, s_\beta))).$$

The Σ_1-formula expressing that the constant term k codes a formula of $\mathcal{L}_{\mathsf{O}}$ is then defined as

$$\mathsf{Fml}_{\mathcal{L}_{\mathsf{O}}}(k) :\leftrightarrow \exists s \exists \alpha \mathsf{FormSeq}_{\mathcal{L}_{\mathsf{O}}}(s, \alpha, k).$$

The notion of being a formula of $\mathcal{L}_{\mathsf{O}}$ beginning with a universal quantifier applied to v_i is formalised as

$$\mathsf{QFml}_{\mathcal{L}_{\mathsf{O}}}(i, k) \leftrightarrow: \ \mathsf{var}(i) \wedge \exists l (\mathsf{Fml}_{\mathcal{L}_{\mathsf{O}}}(l) \wedge k = \mathsf{All}(i, l)).$$

Moreover, it would be useful to have at our disposal a predicate expressing that a closed term of HF codes a complex formula of $\mathcal{L}_{\mathsf{O}}$. This is expressed in $\mathcal{L}_{\mathsf{HF}}$ by the following

Σ_1-formula:

$$\mathsf{CFml}_{\mathcal{L}_\mathsf{O}}(i,k) :\leftrightarrow \mathsf{Fml}_{\mathcal{L}_\mathsf{O}}(k) \wedge \exists l, m, i\ (\mathsf{AppFml}_{\mathcal{L}_\mathsf{O}}(i,l,m,k)).$$

The following Σ_1-formula defines the notion of a formation sequence for a formula k which is built up from atomic formulas by the usual operations $\neg, \rightarrow$ and quantification over variables different from v_i. The formation sequence defined below, in other words, cannot contain entries in which v_i occurs freely and entries in which it is bound.

$$\begin{aligned}\mathsf{SeqQfml}_{\mathcal{L}_\mathsf{O}}(i,s,\alpha,k) :\leftrightarrow \mathsf{var}(i) \wedge \mathsf{SeqEnd}(s,\alpha,k) \wedge (\forall \beta \in \alpha)&\\ \big(\mathsf{AtFml}_{\mathcal{L}_\mathsf{O}}(s_\beta) \vee \mathsf{QFml}_{\mathcal{L}_\mathsf{O}}(i,s_\beta) \vee&\\ (\exists \gamma, \delta \in \beta)(\exists j)(\mathsf{var}(j) \wedge i \neq j \wedge \mathsf{AppFml}_{\mathcal{L}_\mathsf{O}}(j, s_\gamma, s_\delta, s_\beta))\big).&\end{aligned}$$

As we did for terms, we let $\mathsf{SeqQfml}^*_{\mathcal{L}_\mathsf{O}}(i,s,\alpha,k)$ expressing that the formation sequence for the formula k does not have any entry that is not already a subformula of it.

Substitution. The upcoming definitions are mainly pointed at the characterisation of the function $\mathsf{sub}(\ulcorner \varphi \urcorner, i, \ulcorner \mathsf{t} \urcorner) = \ulcorner \varphi(\mathsf{t}) \urcorner$, which formalises in HF the syntactic operation that applies to the formula $\varphi(v)$ of $\mathcal{L}_\mathsf{O}$ and returns the formula $\varphi(\mathsf{t}/v)$. This function plays a crucial role in the case in which O is HF itself for the proof of the diagonal lemma. For our concerns, they are especially needed for the formalisation of the notion of partial substitution, which is in turn required to formalise membership in the set of logical axioms of O.

We first present the Σ_1-formulas expressing (i) that the variable coded by the ordinal i occurs in the term coded by k and (ii) that the variable coded by i does not occur in it:

(i) $\mathsf{varTe}(i,k) :\leftrightarrow \exists s \exists \alpha \mathsf{TermSeq}^*_{\mathcal{L}_\mathsf{O}}(s,\alpha,k) \wedge (\exists \beta \in \alpha)(s_\beta = i)$;

(ii) $\mathsf{nvarTe}(i,k) :\leftrightarrow \exists s \exists \alpha (\mathsf{TermSeq}^*_{\mathcal{L}_\mathsf{O}}(s,\alpha,k) \wedge (\forall \beta \in \alpha)(\mathsf{var}_{\mathcal{L}_\mathsf{O}}(s_\beta) \rightarrow s_\beta \neq i))$.

Similarly, the following Σ_1-formulas expresses that the variable coded by i occur (i) or does not occur (ii) into the atomic formula coded by k:

$$\begin{aligned}
\text{(i) } \mathsf{varAt}(i,k) :\leftrightarrow\ & \exists \beta, \gamma\ \exists k_1, \dots, \exists k_\gamma ((\mathsf{Term}_{\mathcal{L}_\mathsf{O}}(k_1) \wedge \dots \wedge \mathsf{Term}_{\mathcal{L}_\mathsf{O}}(k_\gamma)) \wedge \\
& (\mathsf{varTe}(i,k_1) \vee \dots \vee \mathsf{varTe}(i,k_\gamma)) \wedge k = \dot{R}^\gamma_\beta(k_1,\dots,k_\gamma)) \\
\text{(ii) } \mathsf{nvarAt}(i,k) :\leftrightarrow\ & \exists \beta, \gamma\ \exists k_1, \dots, \exists k_\gamma (\mathsf{Term}_{\mathcal{L}_\mathsf{O}}(k_1) \wedge \dots \wedge \mathsf{Term}_{\mathcal{L}_\mathsf{O}}(k_\gamma) \wedge \\
& \mathsf{nvarTe}(i,k_1) \wedge \dots \wedge \mathsf{nvarTe}(i,k_\gamma) \wedge k = \dot{R}^\gamma_\beta(k_1,\dots,k_\gamma)).
\end{aligned}$$

The notion of being a free variable of a formula of $\mathcal{L}_\mathsf{O}$ is expressed in $\mathcal{L}_\mathsf{HF}$ by the following Σ_1-formula:

$$\mathsf{FreeF}(i,k) :\leftrightarrow \exists s \exists \alpha \Big(\mathsf{SeqQfml}^*_{\mathcal{L}_\mathsf{O}}(i,s,\alpha,k) \wedge (\exists \beta \in \alpha)(\mathsf{AtFml}_{\mathcal{L}_\mathsf{O}}(s_\beta) \wedge \mathsf{varAt}(i,s_\beta))\Big)$$

On the other hand,[11]

$$\begin{aligned}
\mathsf{nFreeF}(i,k) :\leftrightarrow\ & \exists s \exists \alpha \Big(\mathsf{SeqQfml}^*_{\mathcal{L}_\mathsf{O}}(i,s,\alpha,k) \wedge \\
& (\forall \beta \in \alpha)(\mathsf{AtFml}_{\mathcal{L}_\mathsf{O}}(s_\beta) \wedge \mathsf{nvarAt}(i,s_\beta) \vee \mathsf{CFml}_{\mathcal{L}_\mathsf{O}}(k))\Big).
\end{aligned}$$

The following $\mathcal{L}_\mathsf{HF}$-formula expresses that the term coded by t is substitutable for the variable coded by i in the formula coded by k; it formalises the standard recursive definition and it

[11] Cfr. previous footnote.

is clearly Σ_1:

$$\begin{aligned}\mathsf{substTVF}(t,i,k) :\leftrightarrow \mathsf{Term}_{\mathcal{L}_\mathsf{O}}(t) \wedge \exists s \exists \alpha \Big(& \mathsf{SeqQfml}^*_{\mathcal{L}_\mathsf{O}}(i,s,\alpha,k) \wedge \\ & (\forall \beta \in \alpha)\Big(\mathsf{AtFml}_{\mathcal{L}_\mathsf{O}}(s_\beta) \vee (\exists \gamma, \delta \in \beta)(\exists j)\Big(s_\beta = \dot{\neg} s_\gamma \\ & \vee s_\beta = s_\gamma \dot{\to} s_\delta \vee \Big(s_\beta = \mathsf{All}(j, s_\gamma) \wedge \\ & \Big(\mathsf{nfreeF}(i,s_\gamma) \vee (\mathsf{freeF}(i,s_\gamma) \wedge \mathsf{nvarTe}(j,t))\Big)\Big)\Big)\Big)\Big).\end{aligned}$$

The upcoming Σ_1-formula expresses that the formation sequence r of ordinal length α for the $\mathcal{L}_\mathsf{O}$-term coded by the closed $\mathcal{L}_\mathsf{HF}$-term q appearing as the last entry of the sequence has been obtained from the formation sequence s—still of length α—for the term of $\mathcal{L}_\mathsf{O}$ coded by the closed $\mathcal{L}_\mathsf{HF}$-term by replacing all free occurrences of the variable v_i of $\mathcal{L}_\mathsf{O}$ by the term coded by t:

$$\begin{aligned}\mathsf{subseqTVT}(s,r,\alpha,t,i,p,q) :\leftrightarrow \mathsf{var}(i) \wedge \mathsf{SeqEnd}(s,\alpha,p) \wedge \mathsf{SeqEnd}(r,\alpha,q) \wedge & \\ & (\forall \beta \in \alpha)\Big((s_\beta = i \wedge r_\beta = t) \vee \\ & (\mathsf{var}_{\mathcal{L}_\mathsf{O}}(s_\beta) \wedge s_\beta \neq i \wedge r_\beta = s_\beta) \\ & \vee (\mathsf{const}_{\mathcal{L}_\mathsf{O}}(s_\beta) \wedge s_\beta = r_\beta) \vee \\ & \exists \gamma, \delta (\exists \beta_1, \ldots, \exists \beta_\gamma \in \beta)(s_\beta = \dot{f}^\gamma_\delta s_{\beta_1}, \ldots, s_{\beta_\gamma} \wedge \\ & r_\beta = \dot{f}^\gamma_\delta r_{\beta_1}, \ldots, r_{\beta_\gamma})\Big).\end{aligned}$$

The Σ_1-formula of $\mathcal{L}_\mathsf{HF}$ expressing that the $\mathcal{L}_\mathsf{O}$-term coded by the closed $\mathcal{L}_\mathsf{HF}$-term q is obtained by replacing all occurrences of the variable of $\mathcal{L}_\mathsf{O}$ coded by i in the $\mathcal{L}_\mathsf{O}$-term coded by the closed $\mathcal{L}_\mathsf{HF}$-term p is then:

$$\mathsf{subTVT}(t,i,p,q) :\leftrightarrow \exists s \exists r \exists \alpha (\mathsf{subseqTVT}(s,r,\alpha,t,i,p,q))$$

To check that $\mathsf{subTVT}(t, i, p, q)$ has the right properties, we notice that:

Lemma 3.6. For every term $\mathfrak{s}$ of $\mathcal{L}_{\mathsf{O}}$, possibly with v_i free,

$$\mathsf{HF} \vdash \mathsf{subTVT}(\ulcorner \mathsf{t} \urcorner, i, \ulcorner \mathfrak{s} \urcorner, t) \leftrightarrow t = \ulcorner \mathfrak{s}(\mathsf{t}/v_i) \urcorner.$$

The following Σ_1-formula expresses that the formula coded by the constant term l has been obtained by replacing all occurrences of the variable v_i with the term coded by t into the atomic formula coded by k:

$$\begin{aligned}\mathsf{subTVAt}(t, i, k, l) :\leftrightarrow \exists \gamma, \delta, \exists t_1, \dots, t_\gamma \exists z_1, \dots, z_\gamma (&\mathsf{subTVT}(t, i, t_1, z_1) \wedge \dots \wedge \\ &\mathsf{subTVT}(t, i, t_\gamma, z_\gamma) \wedge (k = \underset{\bullet}{R}{}^\gamma_\delta t_1, \dots, t_\gamma \wedge l = \underset{\bullet}{R}{}^\gamma_\delta z_1, \dots, z_\gamma)).\end{aligned}$$

The definition immediately entails:

Lemma 3.7. For φ atomic,

$$\mathsf{HF} \vdash \mathsf{subTVAt}(t, i, \ulcorner \varphi \urcorner, k) \leftrightarrow k = \ulcorner \varphi(\mathsf{t}/v_i) \urcorner.$$

Similarly, we introduce the Σ_1-formula of $\mathcal{L}_{\mathsf{HF}}$ formalising the substitution, in a formation sequence s of length α for a formula coded by k, of all occurrences of the variable coded by i with the term coded by t, so that we obtain the formation sequence r for the formula coded by l, still of ordinal length α:

$$\begin{aligned}\mathsf{subseqTVF}(s,r,\alpha,t,i,k,l) :\leftrightarrow\ & \mathsf{SeqEnd}(s,\alpha,k) \wedge \mathsf{SeqEnd}(r,\alpha,l) \wedge \\ & (\forall\beta \in \alpha)\Big((\mathsf{AtFml}_{\mathcal{L}_\mathsf{O}}(s_\beta) \wedge \mathsf{subTVAt}(t,i,s_\beta,r_\beta)) \vee \\ & ((\exists\gamma,\delta \in \beta)(\exists j)((s_\beta = \dot{\neg} s_\gamma \wedge r_\beta = \dot{\neg} r_\gamma) \vee \\ & (s_\beta = \mathsf{All}(i,s_\gamma) \wedge s_\beta = r_\beta) \vee (s_\beta = s_\gamma \dot{\to} r_\delta \wedge r_\beta = r_\gamma \dot{\to} r_\delta) \vee \\ & (j \neq i \wedge s_\beta = \mathsf{All}(j,s_\gamma) \wedge r_\beta = \mathsf{All}(j,r_\gamma))))\Big).\end{aligned}$$

Now the Σ_1-formula of $\mathcal{L}_\mathsf{HF}$ stating that the formula coded by l is obtained from the formula coded by k by replacing all occurrences the variable coded by the finite ordinal i by the term coded by constant term t is just:

$$\mathsf{subTVF}(t,i,k,l) :\leftrightarrow \exists s,r\exists\alpha(\mathsf{subseqTVF}(s,r,\alpha,t,i,k,l)).$$

As for substitution for terms, we have:

Lemma 3.8. For any formula φ,

$$\mathsf{HF} \vdash \mathsf{subTVF}(\ulcorner \mathsf{t} \urcorner, i, \ulcorner \varphi \urcorner, k) \leftrightarrow k = \ulcorner \varphi(\mathsf{t}/v_i) \urcorner.$$

Finally we can define, by exhibiting a functional formula of $\mathcal{L}_\mathsf{HF}$, the function expressing the result of formally substituting the term t for the variable v_i in the formula φ, all belonging to $\mathcal{L}_\mathsf{O}$:

$$\begin{aligned}\mathsf{sub}(\ulcorner \mathsf{t} \urcorner, i, \ulcorner \varphi \urcorner) = u :\leftrightarrow\ & \Big((\exists! z)(\mathsf{subTVF}(i,\ulcorner \mathsf{t} \urcorner,\ulcorner \varphi \urcorner,z) \wedge \mathsf{subTVF}(i,\ulcorner \mathsf{t} \urcorner,\ulcorner \varphi \urcorner,u))\Big) \vee \\ & \Big(\neg(\exists! z)\mathsf{subTVF}(i,\ulcorner \mathsf{t} \urcorner,\ulcorner \varphi \urcorner,z) \wedge u = 0\Big).\end{aligned} \tag{3.22}$$

It can be now easily verified that $\mathsf{sub}(\ulcorner \mathsf{t} \urcorner, i, \ulcorner \varphi \urcorner) = \ulcorner \varphi(\mathsf{t}/v_i) \urcorner$.

In the next section we shall specify a Hilbert-style calculus in which O will be taken to be formulated. As usual, this choice is motivated by the fact that many metamathematical arguments concerning O are made easier by taking this route. In order to formalise in HF membership of a formula in our chosen axiomatisation of predicate calculus—which is displayed in Table 3.4 below—we require the presence of a Σ_1-predicate expressing that the formula coded by l is the result of replacing *some* occurrences of the variable v_i with the variable v_j in the atomic formula coded by k. This relation is defined in terms of the more basic notion of partial substitution of the occurrences of v_i for v_j in the term coded by t_1 expressed by the following Σ_1-predicate:

$$\begin{aligned}\mathsf{PsubVVT}(j,i,t_1,t_2) :\leftrightarrow\ & \exists s \exists \alpha(\mathsf{TermSeq}^*_{\mathcal{L}_\mathsf{O}}(s,\alpha,t_1) \wedge \mathsf{TermSeq}^*_{\mathcal{L}_\mathsf{O}}(r,\alpha,t_2)\\ & \wedge (\forall \beta \in \alpha)(s_\beta = r_\beta \vee (s_\beta = i \wedge r_\beta = j))).\end{aligned}$$

Therefore we can define:

$$\begin{aligned}\mathsf{PsubVVAt}(i,j,k,l) :\leftrightarrow\ & \exists \gamma, \delta\ \exists t_1,\dots,t_n(\mathsf{Term}_{\mathcal{L}_\mathsf{O}}(t_1) \wedge \dots \wedge \mathsf{Term}_{\mathcal{L}_\mathsf{O}}(t_\gamma) \wedge\\ & k = \underset{\cdot}{R}^\gamma_\delta t_1,\dots,t_\gamma \wedge \exists t^*_1,\dots,t^*_\gamma(\mathsf{PsubVVT}(j,i,t_1,t^*_1) \wedge \dots \wedge\\ & \mathsf{PsubVVT}(j,i,t_\gamma,t^*_\gamma)) \wedge l = \underset{\cdot}{R}^\gamma_\delta t^*_1,\dots,t^*_n).\end{aligned}$$

Provability. In order to formalise the notion of being a proof in our object theory O, we need to specify the logical system in which O is formulated. As it was just emphasised, it is convenient to assume that O is developed in a Hilbert-style calculus. Table 3.4 displays the logical axioms of O. We might safely assume that the identity symbol of $\mathcal{L}_\mathsf{O}$, which we denote with $\doteq$ to distinguish it from the identity symbol of HF, is R^2_0, so that $\ulcorner \doteq \urcorner = (\ulcorner R \urcorner, (0,2))$. The set of nonlogical axioms of O will obviously depend on our choice of O, and we can postpone their formalisation, although it should be clear at this stage that we can construct a predicate expressing membership in the set of nonlogical axioms of O which is no worse than $I\Sigma_1$.

L1	$\varphi \to (\varphi \to \psi)$;
L2	$\varphi \to (\psi \to \chi) \to (\varphi \to \psi) \to (\varphi \to \chi)$;
L3	$(\neg\varphi \to \neg\psi) \to (\psi \to \varphi)$;
L4	$\forall v_i \varphi \to \varphi(\mathsf{t}/v_i)$ where t is substitutable for v_i in φ;
L5	$\forall v_i(\varphi \to \psi) \to (\varphi \to \forall v_i \psi)$ with v_i not free in φ;
L6	$\forall v_i(v_i \dot{=} v_i)$
L7	$v_i \dot{=} v_j \to (\varphi \to \psi)$ where φ is atomic and ψ is obtained from φ by substituting *some* occurrences of v_i with v_j.

Table 3.4: Logical Axioms of O.

Moreover, if O has L1- L7 as logical axioms, then it will involve two rules of inference, modus ponens and generalisation:

MP: from φ and $\varphi \to \psi$, infer ψ;

Gen: from φ, infer $\forall v_j \varphi$.

We can now express the Σ_1-formula of $\mathcal{L}_{\mathsf{HF}}$ stating that x is the code of a logical axiom of

O:[12]

$$
\begin{aligned}
\mathsf{AxLo}_O(k) :\leftrightarrow \exists l, m, n(&\mathsf{Fml}_{\mathcal{L}_\mathsf{O}}(l) \wedge \mathsf{Fml}_{\mathcal{L}_\mathsf{O}}(m)\mathsf{Fml}_{\mathcal{L}_\mathsf{O}}(n) \wedge \\
&k = \mathsf{Imp}(l, \mathsf{Imp}(m, l)) \vee \\
&k = \mathsf{Imp}(\mathsf{Imp}(l, \mathsf{Imp}(m, n)), \mathsf{Imp}(\mathsf{Imp}(l, m), \mathsf{Imp}(l, n))) \vee \\
&k = \mathsf{Imp}(\mathsf{Imp}(\mathsf{Neg}(l), \mathsf{Neg}(m)), \mathsf{Imp}(m, l)) \vee \\
&(\mathsf{substTVF}(\ulcorner \mathsf{t} \urcorner, i, m) \wedge \mathsf{subTVF}(\ulcorner \mathsf{t} \urcorner, i, m, n) \wedge k = \mathsf{Imp}(\mathsf{All}(i, m), n)) \vee \\
&(\mathsf{nFreeF}(i, l) \wedge k = \mathsf{Imp}(\mathsf{All}(i, \mathsf{Imp}(l, m)), \mathsf{Imp}(l, \mathsf{All}(i, m)))) \vee \\
&(k = \mathsf{All}(i, \mathsf{Eq}(i, i))) \vee \\
&(\mathsf{PsubVVAt}(i, j, l, m) \wedge k = \mathsf{Imp}(\mathsf{Eq}(i, j), \mathsf{Imp}(l, m)))).
\end{aligned}
$$

Since we did not specify a particular axiomatization of O, we assume that the (no-worse than Σ_1) formula $\mathsf{Ax}_\mathsf{O}(k)$ expresses that the formula coded by k is a nonlogical axiom of O.

The following Σ_1-formulas express that a formula coded by n is obtained from a formula coded by l and the formula coded by m by application of modus ponens and that m is obtained from l by application of generalisation:

$$
\begin{aligned}
\mathsf{MP}(l, m, n) &:\leftrightarrow m = \mathsf{Imp}(l, n); \\
\mathsf{Gen}(l, m) &:\leftrightarrow \exists i(\mathsf{Fml}_{\mathcal{L}_\mathsf{O}}(l) \wedge \mathsf{Fml}_{\mathcal{L}_\mathsf{O}}(m) \wedge m = \mathsf{All}(i, l)).
\end{aligned}
$$

We now have at our disposal all the components that we need to express within HF that the constant term s is a proof of ordinal length α of a formula coded by x:

$$
\begin{aligned}
\mathsf{B}_O(s, \alpha, k) :\leftrightarrow \mathsf{SeqEnd}(s, \alpha, k) \wedge (\forall \beta \in \alpha)(&\mathsf{AxLo}_O(s_\beta)) \vee \mathsf{Ax}_O(s_\beta) \vee \\
&(\exists \gamma, \delta \in m)(\mathsf{MP}(s_\gamma, s_\delta, s_\beta) \vee (\mathsf{Gen}(s_\gamma, s_\beta))).
\end{aligned} \tag{3.23}
$$

[12] We opted for a somewhat abbreviated definition of L1-L3 as their formalization is straightforward.

We can finally introduce the Σ_1-formula

$$\mathsf{Bew}_O(k) :\leftrightarrow \exists s \exists \alpha (\mathsf{B}_\mathsf{O}(s, \alpha, k)),$$

expressing that the formula coded by x is a theorem of O.

The following result holds under the innocuous assumption that HF is 'sound' (actually Σ_1-soundness would suffice), in the sense that we have a clear access to the standard model $\mathcal{H}$ of HF. Although we are assuming that the language of O and the language of HF are disjoint, we can in fact prove

Proposition 3.9 (Provability Condition*)**.** For every formula φ of $\mathcal{L}_\mathsf{O}$,

$$O \vdash \varphi \Leftrightarrow \mathsf{HF} \vdash \mathsf{Bew}_O(\ulcorner \varphi \urcorner)$$

Proof. ($\Rightarrow$): if $\mathsf{O} \vdash \varphi$, then there is a finite sequence $\langle \varphi_\beta : \beta < \alpha \rangle$ in which every φ_β is a logical or nonlogical axiom of O or it is obtained from previous formulas in the sequence by application of MP or Gen. We can thus construct the sequence

$$s = \{(\beta, \ulcorner \varphi_\beta \urcorner) : \beta < \alpha\}$$

so that $s \in \mathbb{H}$, that is s belongs to the domain of the standard model of HF.

Thus simply

there are an $s \in \mathbb{H}$ and an ordinal α such that	$\mathcal{H} \vDash \mathsf{B}_\mathsf{O}(s, \alpha, \ulcorner \varphi \urcorner)$	
	$\mathcal{H} \vDash \mathsf{Bew}_\mathsf{O}(\ulcorner \varphi \urcorner)$	truth in $\mathcal{H}$
	$\mathsf{HF} \vdash \mathsf{Bew}_\mathsf{O}(\ulcorner \varphi \urcorner)$	Prop. 3.3.

($\Leftarrow$): if $\mathsf{HF} \vdash \mathsf{Bew}_\mathsf{O}(\ulcorner \varphi \urcorner)$, then $\mathcal{H} \vDash \exists s \exists \alpha \mathsf{B}_\mathsf{O}(s, \alpha, \ulcorner \varphi \urcorner)$, so there is a sequence $s \in \mathbb{G}$ such that $s_{\overline{\alpha}}$ is $\ulcorner \varphi \urcorner$ and $s = \langle s_\beta : \beta < \alpha \rangle$, where each s_β with $\beta \in \alpha$ is an axiom of O or obtained from earlier values by MP or Gen, thus $\mathsf{O} \vdash \varphi$. *qed.*

Proposition 3.9 completes the route started in Section §3.1. Our informal characterisation of the syntax of the language $\mathcal{L}_{\mathsf{O}}$ and of syntactic and logical notions concerning O has now been fully captured by HF. In the next section we shall consider some advantages that HF displays with respect to the usual strategies.

3.4 Why HF.

The previous chapter concluded with the outline of a project: we wanted to investigate an axiomatic framework which could somewhat mimic the original Tarskian picture of the metatheory, in which the theory dealing with the syntax of the object theory was fixed at the outset, independently from the choice of the object theory itself. In this way, we argued, it is possible to recover a portion of the generality proper of the Tarskian method and to clarify the interactions between schemata of the theory of truth and schemata of the object theory that are blurred by the usual construction. As we shall see in the next chapter, the target is thus an axiomatisation of the truth predicate in which the domain of 'syntactic' objects—that is objects to which truth is ascribed—and the domain of 'mathematical' objects, namely the entities in the domain of discourse of the object theory, are clearly distinguished.

To present arguments in some detail, however, it is necessary to fix a theory which acts as theory of syntax. This is also the path chosen by [Heck 2009] and [Leigh&Nicolai 2013], who investigate systems akin to the ones that will be introduced in the next chapter by choosing Peano Arithmetic or fragments of it ($I\Sigma_1$ in Heck's case) as the theories in which the syntax of O is formalised. The informal account of the syntax of a first order language given in §3.1 is usually assumed also when PA acts as its formalisation: we thus found HF as a much more natural environment to capture this informal characterisation. HF in fact inherently captures the tree-like structure presupposed by the understanding of the grammar of formal languages akin to §3.1. This also renders the present work totally self-contained: unlike the case of PA, all the 'syntactic' notions that are introduced do not rely on hidden nontrivial results, such as the ones leading to the description of the notion of finite sequence in arithmetical context.

Besides an obvious glimpse of pedantry, the possibility of writing down in full detail the representing formulas for syntactic and logical notions concerning the object theory and its language has the good effect of ruling out intensionally incorrect representations of those notions, that is formulas that define syntactic notions without fully expressing their 'meaning'.[13] This, however, marks no difference with respect to the arithmetical context: also in Peano Arithmetic and its fragment those nonstandard representations are ruled out if one lists representing formulas for syntactic and logical notions and the defining equations for syntactic operations. HF simply offers a much more comfortable environment to carry out the task and to verify that the descriptions obtained satisfy the intended syntactic properties.

The naturality displayed by the descriptions carried out above has some impact also in the proofs of Gödel's Incompleteness Theorems. A surprisingly detailed proof of both theorems, especially the second, can be carried out in HF by taking HF itself as the theory O in the development of the syntax above. For full proofs, we refer to Świerczkowski 's paper. We outline the strategy and the main ingredients of the proofs in the Appendix.

The advantages of presenting Gödel's Incompleteness theorems in HF are also pedagogical: the presentation of the proofs of both theorems is carried out step-by-step, assuming only a relatively little knowledge of HF. Moreover, it makes possible to formally verify of Gödel's Second theorem, a task that was never achieved for the arithmetical setting.[14]

---

[13] Cfr. [Feferman 1960].

[14] The result was announced on FOM Mailing List by Larry Paulson, Cambridge University. The proofs were implemented in the software Isabelle/HOL.

3.5 Appendix: Incompleteness Theorems in HF

Let us assume now that $\mathcal{L}_{\mathsf{O}}$ is $\mathcal{L}_{\mathsf{HF}}$ itself. Let $\ulcorner 0 \urcorner = 0$, $\ulcorner \triangleleft \urcorner = (0, 0)$ and $\ulcorner \in \urcorner = (0, 0, 0)$. Thus there are no further function and relation symbols. Codes of logical constants are chosen among distinct members of $\mathbb{G}$ different from the ones for $0, \triangleleft, \in$. Moreover, we need to introduce the formalisation of some more specific syntactic categories for $\mathcal{L}_{\mathsf{HF}}$, beginning with the notion of closed terms of $\mathcal{L}_{\mathsf{HF}}$, expressed by the Σ_1-formula

$$\mathsf{Cterm}(t) :\leftrightarrow \exists\alpha\exists k(\mathsf{SeqEnd}(s, \alpha, t) \wedge (\forall\beta \in \alpha)(s_\beta = 0 \vee (\exists\gamma, \delta \in \beta)(s_\beta = (\ulcorner \triangleleft \urcorner, s_\gamma, s_\delta)))).$$

First Incompleteness Theorem. The function $h(x)$ defined by recursion on rank in Section 3.2—immediately after Proposition 3.5—has the property that for any element t of $\mathbb{G}$,

$$h(t) = \ulcorner t \urcorner \tag{3.24}$$

Let us now consider the function, for some fixed i,

$$d(k) := \mathsf{sub}(i, h(k), k) \tag{3.25}$$

Where $\mathsf{sub}(k, l, m)$ is as in (3.22). This function exists and yields, for any φ with only v_i free,

$$d(\ulcorner \varphi \urcorner) = \ulcorner \varphi(\ulcorner \varphi \urcorner) \urcorner. \tag{3.26}$$

The *diagonal lemma* is thus readily obtained: given any formula $\psi(v_i)$, let

$$\theta(v_i) :\leftrightarrow \exists k(k = d(v_i) \wedge \psi(k))$$

Thus

$$\mathsf{HF} \vdash \theta(v_i) \leftrightarrow \psi(d(v_i))$$

and

$$\mathsf{HF} \vdash \theta(\ulcorner\theta\urcorner) \leftrightarrow \psi(\ulcorner(\ulcorner\theta\urcorner)\urcorner).$$

Therefore $\theta(\ulcorner\theta\urcorner)$ is the required fixed point.

Theorem 3.1 (Gödel's First Incompleteness Theorem)**.** There is a sentence γ such that $\mathsf{HF} \nvdash \gamma$ and $\mathsf{HF} \nvdash \neg\gamma$.

Proof. Consider $\neg\mathsf{Bew}_{\mathsf{O}}(v_i)$ a let γ be its fixed point obtained by the diagonal method.

If $\mathsf{HF} \vdash \gamma$, then $\mathsf{HF} \vdash \neg\mathsf{Bew}_{\mathsf{HF}}(\ulcorner\gamma\urcorner)$ by the diagonal lemma. But also $\mathsf{HF} \vdash \mathsf{Bew}_{\mathsf{HF}}(\ulcorner\gamma\urcorner)$, by Proposition 3.9, contradicting the consistency of HF.

If $\mathsf{HF} \vdash \neg\gamma$, then $\mathsf{HF} \vdash \mathsf{Bew}_{\mathsf{HF}}(\ulcorner\gamma\urcorner)$ and $\mathsf{HF} \vdash \gamma$ by Proposition 3.9. Again a contradiction.

qed.

As a consequence, we can provide an argument for the truth of that Gödel sentence γ in the standard model of HF: assume not, then $\mathcal{H} \vDash \neg\gamma$. By the definition of γ, this yields $\mathcal{H} \vDash \mathsf{Bew}_{\mathsf{HF}}(\ulcorner\gamma\urcorner)$. By Propositions 3.3 and 3.9, we have $\mathsf{HF} \vdash \gamma$. Thus $\mathcal{H} \vDash \gamma$ after all.

Second Incompleteness Theorem. Let the consistency statement for HF in $\mathcal{L}_{\mathsf{HF}}$ be

$$\mathsf{Con}_{\mathsf{HF}} :\leftrightarrow \neg\mathsf{Bew}_{\mathsf{HF}}(\ulcorner 0 \in 0\urcorner).$$

In order to obtain the Second Incompleteness Theorem, it suffices to show the implication $\mathsf{Con}_{\mathsf{HF}} \to \sigma$, where σ is the Gödel sentence of Proposition 3.1.

The claim is obtained via two important lemmata. Let $i(x)$ as in the example immediately after Proposition 3.7; moreover, the coding presented in Section 2.3. has to be slightly modified to be able to code expressions of $\mathcal{L}_{\mathsf{HF}}$ via *terms*, and not only closed terms of $\mathcal{L}_{\mathsf{HF}}$. In particular, a coding $\lceil . \rceil$ is defined so that given a $\mathcal{L}_{\mathsf{HF}}$-formula $\varphi(v_1, \ldots, v_n)$, then the free variables of $\lceil \varphi \rceil$ are exactly $v_1, \ldots, v_n$. $\lceil \varphi \rceil(\mathrm{t}_1, \ldots, \mathrm{t}_n)$ denotes the result of substituting $\lceil . \rceil$ has also the property that, if φ is a sentence, then $\lceil \varphi \rceil = \ulcorner \varphi \urcorner$.[15]

There are two main lemmata: we limit ourselves to state them. For full proofs, we refer to [Świerczkowski 2003].

The first important lemma is:

Lemma 3.9. If $\mathsf{HF} \vdash \varphi(v_1, \ldots, v_n) \rightarrow \psi(v_1, \ldots, v_n)$, then

$$\mathsf{HF} \vdash \mathsf{Bew}(\lceil \varphi \rceil(i(v_1), \ldots, i(v_n)) \rightarrow \mathsf{Bew}(\lceil \psi \rceil(i(v_1), \ldots, i(v_n)))).$$

The second lemma is the equivalent of one of the well-known Löb's conditions, and it is obtained by induction on the length of the formula involved:

Lemma 3.10. For every Σ_1-formula, $\varphi(v_1, \ldots, v_n)$ with only the free variables displayed, then

$$\mathsf{HF} \vdash \varphi(v_1, \ldots, v_n) \rightarrow \mathsf{Bew}_{\mathsf{HF}}\lceil \varphi \rceil(i(v_1), \ldots, i(v_n)).$$

Now from the last lemma, we have

$$\mathsf{HF} \vdash \mathsf{Bew}_{\mathsf{HF}}(\ulcorner \gamma \urcorner) \rightarrow \mathsf{Bew}_{\mathsf{HF}}(\ulcorner \mathsf{Bew}(\ulcorner \gamma \urcorner) \urcorner) \tag{3.27}$$

Since $\gamma \leftrightarrow \neg\mathsf{Bew}_{\mathsf{HF}}(\ulcorner \gamma \urcorner)$, we have by Lemma 3.9,

$$\mathsf{HF} \vdash \mathsf{Bew}_{\mathsf{HF}}(\ulcorner \gamma \urcorner) \rightarrow \mathsf{Bew}_{\mathsf{HF}}(\ulcorner \mathsf{Bew}_{\mathsf{HF}}(\ulcorner \gamma \urcorner) \wedge \neg\mathsf{Bew}_{\mathsf{HF}}(\ulcorner \gamma \urcorner) \urcorner) \tag{3.28}$$

[15] For further details, see [Świerczkowski 2003, §7], in particular, the definition of the coding $\lceil . \rceil$ has to go through an important intermediate step that allows an inductive definition.

But $\mathsf{Bew}(\ulcorner\gamma\urcorner) \wedge \neg\mathsf{Bew}(\ulcorner\gamma\urcorner)$ implies $0 \in 0$, thus again by Lemma 3.9,

$$\mathsf{HF} \vdash \mathsf{Bew}_{\mathsf{HF}}(\ulcorner\gamma\urcorner) \to \mathsf{Bew}_{\mathsf{HF}}(\ulcorner 0 \in 0\urcorner). \tag{3.29}$$

The last line yields $\mathsf{Con}_{\mathsf{HF}} \to \gamma$. Thus:

Theorem 3.2 (Second Incompleteness Theorem)**.** If HF is consistent, then

$$\mathsf{HF} \nvdash \mathsf{Con}_{\mathsf{HF}}.$$

4 Theories of Truth with 'Disentangled' Syntax

In Chapter 2 we have tried to focus the reader's attention on the possibility of an alternative way of constructing (axiomatic) theories of truth in which syntactic and logical notions concerning the object theory O are formalised in a theory disjoint from O itself. We isolated historical and more theoretical motivations: in particular, the generality of Tarski's metatheory and the peculiar role that axiom schemata play in the usual framework suggested us to explore an unconventional pattern.

The picture underlying the main construction that will be shortly introduced has recently attracted new attention among truth-theorists (see [Heck 2009] and [Halbach 2011]) but the general idea, as we have seen, can be traced back to Tarski's seminal paper. In the previous chapter we have introduced the theory that will talk about the objects to which truth is ascribed by our theory of truth. In the present our target will be to introduce the theory $\mathsf{CTD}[\mathsf{O}]$, standing for compositional truth with 'disentangled' syntax.

After introducing the many-sorted language $\mathcal{L}_{\mathsf{TD}}$, featuring quantification over the distinct domains of objects of truth and of 'mathematical' objects in the sense of our object theory, we will consider some weaker axiomatisations of the resulting 'disentangled' semantics. Once the full $\mathsf{CTD}[\mathsf{O}]$ is presented, we will evaluate the status of explicit assertions of soundness and consistency for our object theory O formulated in $\mathcal{L}_{\mathsf{TD}}$ and in the language ($\mathcal{L}_{\mathsf{HF}}$) of our syntax theory respectively; we will then compare these 'syntactic' consistency statements to the 'usual' ones, formalisable in the language of O itself. To this purpose, it will also be interesting to consider some extensions of $\mathsf{CTD}[\mathsf{O}]$: in the present chapter we will be mainly concerned with the setting in which the schemata of O, if present, are extended to the entire vocabulary of $\mathcal{L}_{\mathsf{O}}$.

4.1 The language $\mathcal{L}_{\mathsf{TD}}$

We present the three-sorted language $\mathcal{L}_{\mathsf{TD}}$. We start with a set of sorts $\mathsf{I} = \{o, s, sq\}$. The first sort o labels variables ranging over the domain of the object theory O, s labels variables ranging over hereditarily finite sets, while sq labels variables ranging over a third domain of variable assignments, 'mixed' objects of both syntactic and mathematical nature as they associate to syntactic objects elements of the domain of discourse of O.

For clarity we will employ elements of I also to label nonlogical constants of $\mathcal{L}_{\mathsf{TD}}$ when necessary, so that, for instance, $\mathsf{Sent}^s_{\mathcal{L}_{\mathsf{O}}}(t)$ expresses in HF that the expression coded by the $\mathcal{L}_{\mathsf{HF}}$-term t is a sentence of $\mathcal{L}_{\mathsf{O}}$. Equivalently—as the semantics is identical—we might consider a single-sorted language expanding $\mathcal{L}_{\mathsf{O}}$ with three unary predicate symbols P_o, P_s and P_{sq} relativising variables to range over the domain of discourse $|\mathcal{M}|$ of a model $\mathcal{M}$ of O, to the domain of a model $\mathfrak{H}$ of HF, and to a class $\mathcal{S}$ of objects encoding sequences in $|\mathcal{M}|$. We choose, for the sake of readability, to define $\mathcal{L}_{\mathsf{TD}}$ according to the first of these two equivalent formulations.

As a consequence, the set of logical symbols of $\mathcal{L}_{\mathsf{TD}}$, besides $\neg$, $\rightarrow$ and $\forall$, will include a disjoint family of nonempty sets

$$V = V_o \sqcup V_s \sqcup V_{sq} \tag{4.1}$$

of countably many variables of each sort. Consistently with the notation already introduced in the previous chapter, we denote with $v^o_1, v^o_2, v^o_3, \ldots$ and occasionally with $x, y, z, \ldots$ elements of V_o, with $v_1^s, v_2^s, v_3^s, \ldots$—abbreviated by $i, j, k, \ldots$—variables ranging over hereditarily finite sets, and with $a, b, c\ldots$ variables in V_{sq} abbreviating the official enumeration $v_1^{sq}, v_2^{sq}, v_3^{sq}, \ldots$. Each sort of variables might be expanded with indexes if necessary. Logical constants may include also binary equality symbols $=_o$, $=_s$ and $=_{sq}$. For the sake of readability, we will omit them.

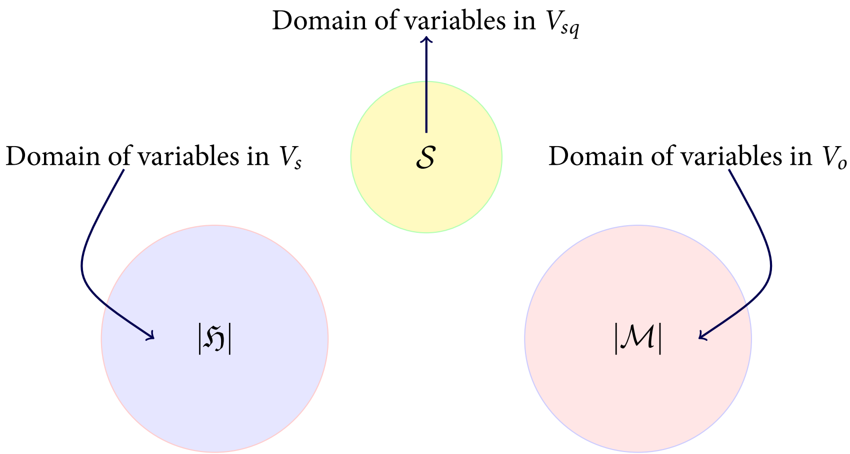

Figure 4.1: Subdomains of discourse of $\mathcal{L}_{\mathsf{TD}}$.

Nonlogical constants of $\mathcal{L}_{\mathsf{TD}}$ will comprehend o, $\in$ and $\lhd$ belonging to $\mathcal{L}_{\mathsf{HF}}$, the nonlogical constants of the object language O, and some nonlogical symbols of 'mixed nature', which are required for the formulation of the theory of truth. In particular, since $\mathcal{L}_{\mathsf{O}}$ may contain terms, our formulation of $\mathcal{L}_{\mathsf{TD}}$ involves two function symbols $.(.)$ and D, both of type $(sq, s) \to o$, which apply to a sequence and to a syntactic object and yield a mathematical object, and a binary predicate Sat of sort (sq, s).

In order to simplify the notation, we recall that $\mathfrak{r}, \mathfrak{s}, \mathfrak{t}, \ldots$ range over elements terms of $\mathcal{L}_{\mathsf{O}}$ and $\varphi, \psi, \chi, \ldots$ over formulas of $\mathcal{L}_{\mathsf{O}}$. We will in fact quantify directly over codes of terms and formulas, so that

- $\forall \ulcorner \varphi \urcorner$ is short for $\forall k(\mathsf{Fml}^s_{\mathcal{L}_O}(k) \to \ldots)$;

- $\forall^{\ulcorner}\mathsf{t}^{\urcorner}$ is short for $\forall k(\mathsf{Term}^{s}_{\mathcal{L}_O}(k) \to ...)$;
- $\forall^{\ulcorner}\varphi(v_i)^{\urcorner}$ is short for $\forall k \forall i(\mathsf{Sent}^{s}_{\mathcal{L}_O}(\underset{\cdot}{\forall} v_i k) \to ...)$.

We will employ capital, capital letters $A, B, C...$ to range over formulas of $\mathcal{L}_{\mathsf{TD}}$. Finally, according to our official notation for ordinals, we should index symbols of $\mathcal{L}_{\mathsf{O}}$ and code variables of $\mathcal{L}_{\mathsf{O}}$ by resorting to variables denoting ordinals $\alpha, \beta, \gamma, ...$ For simplicity, however, indexes will still be denoted by $i, j, k, l, m...$, but it should be kept in mind, as it will be clear from the context, that they are ordinals in the sense of HF.

4.2 Disquotational Truth and Material Adequacy

An axiomatisation of the truth predicate based solely on the T-sentences,[1] that is a suitable formalisation of the schema

(4.2) ‘p’ is a true sentence if and only if p,

is already discussed in Tarski’s seminal paper (cfr. [Tarski 1956a, p.256]) and dismissed there as ‘highly incomplete’ according to the adequacy criteria set up earlier in the paper.[2] Proponents of the deflationary standpoint about truth have considerably revived the interest for such axiomatisations: deflationism seems in fact to be deeply tied with an axiomatic approach to truth (see for instance [Horwich 1998] and [Horsten 2011]). Moreover, according to some versions of deflationism, the disquotation schema fixes the extension of the truth predicate: there is nothing more to learn about truth than what Tarski-biconditionals tell us; truth is ‘at bottom disquotational’.[3]

It is not entirely clear how the biconditional proper of the T-schema has to be understood: the equivalence between $T\overline{\sigma}$ and σ, where σ is a metavariable for sentences of English, T

[1] We will also use the locution ‘disquotation scheme’ to refer to the T-schema.
[2] However, see [Halbach 2011] and the discussion below to question Tarki’s adequacy requirements.
[3] Cfr. [Field 1994].

the truth predicate and $\ulcorner\urcorner$ a name forming device, has sometimes been explained in terms of analiticity or also, in a more Quinean spirit, in terms of cognitive equivalence.[4]

In what follows we will set aside philosophical issues related to the T-schema, and focus mainly on the axiomatisations of the truth predicate based on it. In particular, we will introduce the theory $\mathsf{TBD}[O]$, axiomatising the notion of truth for the object theory O in a setting with 'disentangled' syntax.

Let us first define an intermediate system.

Definition 4.1 (HFD)**.** HFD is the system formulated in the language $\mathcal{L}_{\mathsf{TD}}$ and whose axioms are $\mathsf{HF1}$, $\mathsf{HF2}$ and the axiom schema

$$\mathsf{HF3}^* \qquad A(0) \wedge \forall k, l(A(k) \wedge A(l) \rightarrow A(k \lhd l)) \rightarrow \forall k\, A(k),$$

where $A(k)$ is an arbitrary formula of $\mathcal{L}_{\mathsf{TD}}$ and k is of sort s.

HFD is conservative over HF, as the new vocabulary is idling in the absence of suitable axioms.

$\mathsf{TBD[O]}$ will obviously contain the axioms of O. No restriction at this stage is required, O can be any first-order theory. Moreover, $\mathsf{TBD[O]}$ will contain the axioms of HFD defined above, and the axiom

$$\mathsf{SQ} \qquad \forall a \forall x \forall j \exists b(\forall i(i \neq j \rightarrow a(i) = b(i)) \wedge b(j) = x),$$

which tells us that, given any sequence a of mathematical objects, we can always construct a sequence b which is exactly like a except for one of its entries, to which we can assign an arbitrary value. Moreover, it follows logically from the axioms that the domain of sequences is nonempty.

This axiom is reminiscent of axioms (III)1-2 in [Craig & Vaught 1958] in which, by means

[4][Halbach 2001b] tries to provide a precise account of analiticity.

of a theory of sequences and of axioms for satisfaction, a finitely axiomatised extension for any recursively axiomatised theory is constructed.[5] It should also be noticed that, standing this formulation, SQ does not stipulate anything about the length of our special sequences. We observe, *en passant*, that HFD + SQ is also conservative over HF as any model of HF can be expanded to a model of HFD + SQ.

Finally, TBD[O] will contain the T-schema for $\mathcal{L}_{\mathsf{O}}$ formulated in $\mathcal{L}_{\mathsf{TD}}$, that is for each sentence σ in $\mathcal{L}_{\mathsf{O}}$, we have

$$\mathsf{Tbd} \qquad \forall a\, \mathsf{Sat}(a, \ulcorner\sigma\urcorner) \leftrightarrow \sigma.$$

We recall that $\ulcorner\sigma\urcorner$ is a (closed) term of $\mathcal{L}_{\mathsf{HF}}$ and then its denotation lives in the realm of hereditarily finite sets.

We can thus define TBD[O]:

Definition 4.2 (TBD[O]). TBD[O] is the first-order theory in $\mathcal{L}_{\mathsf{TD}}$ whose axioms are the axioms of O, HF1, HF2, HF3*, SQ and the schema Tbd for each $\mathcal{L}_{\mathsf{O}}$-sentence.

Let

$$\ulcorner\varphi\urcorner[\ulcorner\mathrm{t}\urcorner/i] := \mathsf{sub}(\ulcorner\mathrm{t}\urcorner, i, \ulcorner\varphi\urcorner);$$

that is, $\ulcorner\varphi\urcorner[\ulcorner\mathrm{t}\urcorner/i]$ denotes the finite set coding the result of formally substituting the term t of $\mathcal{L}_{\mathsf{O}}$ for the variable of sort o coded by i in the $\mathcal{L}_{\mathsf{O}}$-formula φ. The theory UTBD[O] is obtained by replacing the schema Tbd in the definition of TBD[O] with the schema, for every formula $\varphi(v_i)$ of $\mathcal{L}_{\mathsf{O}}$ with only the free variable displayed,

$$\mathsf{Utbd} \qquad \forall a \forall \ulcorner\mathrm{t}\urcorner(\mathsf{Sat}(a, \ulcorner\varphi\urcorner[\ulcorner\mathrm{t}\urcorner/i]) \leftrightarrow \varphi(D(a, \ulcorner\mathrm{t}\urcorner));$$

and by adding to TBD[O] the axioms CTDv, CTDc and CTDf from Definition 4.3 below. It

[5] Actually, the axiomatisation presented in [Heck 2009] and that will be generalised in what we will call CTD[O] can be seen as a reformulation of Craig's and Vaught's setting.

is then clear that $\mathsf{TBD[O]}$ is a subsystem of $\mathsf{UTBD[O]}$.

The status of $\mathsf{TBD[O]}$ appears to be relevant for a discussion of the adequacy requirements imposed by Tarski in the *Wahrheitsbegriff*. At the beginning of section 3 of his founding paper, before providing a definition in the metalanguage of the notion of true sentence for the language of the Calculus of Classes, Tarski introduces the celebrated *Convention T*:

> A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an *adequate definition of truth* if it has the following consequences:
>
> (α) all sentences which are obtained from the expression '$x \in Tr$ if and only if p' by substituting for the symbol 'x' a structural descriptive name of any sentence of the language in question and for the symbol 'p' the expression which forms the translation of this sentence into the metalanguage;
>
> (β) the sentence 'for any x, if $x \in Tr$ then $x \in S$' (in other words '$Tr \subseteq S$'). ([**?**, pp. 187f])

In the quotation, S is just the set of sentences of the object language.

We have already discussed the notion of metatheory employed by Tarski, and there is no need to remark that we will mainly be concerned with axiomatic theories of truth in which the object theory is included in the 'metatheory'. Therefore the semantic notion of translation employed in the *Convention T*, although somewhat problematic, is not among our primary concerns.

Convention T tells us that, in order to meet Tarski's adequacy criteria, the metatheory must prove the T-sentences for the object-language. In section 5 of the *Wahrheitsbegriff*, Tarski discusses the possibility of axiomatising the truth predicate over the metatheory by employing the equivalences from part (α) of *Convention T*.[6]

> The idea naturally suggests itself of setting up semantics as a special deductive science with a system of morphology as its logical substructure. For this purpose

[6] We will call them instances of the T-schema, or of the disquotation schema interchangeably.

> it would be necessary to introduce into morphology a given semantical notion as an undefined concept and to establish its fundamental properties by means of axioms. ([Tarski 1956a, p. 255])

The passage is quite relevant for our general purposes, as it will explained shortly. Tarski focuses on the extension of the 'morphology' of an object theory O obtained by adding to it all instances of the T-schema. The consistency of the resulting theory can be proved by noticing that we can always replace the truth predicate, in each of its finite subtheories, by a defined truth predicate, so that the (finitely many) instances of the disquotation schema proper of each subtheory are turned into provable sentences on the 'morphology' of O, that is of the metatheory.[7] Thus if the theory with the T-schema is inconsistent, the inconsistency can be replicated in a finite class of sentences of the 'morphology' of O, so the theory considered must be consistent after all if the 'morphology' of O is. The same proof, with little modification, also yields a proof of conservativity of the theory of truth based in the equivalences from (α) over the 'morphology' of O.

In contemporary discussions, it would seem natural to think of theories in the style of TB,[8] in which the T-scheme is added to Peano Arithmetic formulated in the language $\mathcal{L}_{\mathsf{A}} \cup \{T\}$, where $\mathcal{L}_{\mathsf{A}}$ is the language of arithmetic, as the axiomatic theories of truth that better capture the spirit of Tarski's proposal. The similarities are obvious, but if our understanding of the structure of the Tarskian metatheory—discussed in Chapter 2—goes in the right direction, it seems that $\mathsf{TBD}[\mathsf{O}]$ is even closer to the setting contemplated by Tarski in his Theorem III.[9] Tarski is in fact interested in the axiomatisation of the truth predicate over the metatheory. But if the metatheory has the structure of a two-sorted theory in which syntactic and mathematical objects belong to different domains, then it is natural to consider $\mathsf{TBD}[\mathsf{O}]$ as a plausible alternative. In other words, if the goal of the axiomatisation of the truth predicate is to prove the equivalences contained in (α), then it seems that the inclusion

[7] Which is actually quite similar to the theory that we will call $\mathsf{S{\cdot}O}$, .
[8] Cfr. §2.3.2.
[9] Cfr. [Tarski 1956a, p. 256].

of the theory of syntax into the object theory is not required. *According to* Convention T, *in fact,* $\mathsf{TDB[O]}$ *is a materially adequate theory of truth for* $\mathcal{L}_O$.

[Halbach 2011] casts some doubts concerning the compatibility of Tarski's intuitive adequacy requirements and the ones actually expressed by *Convention T*. The theory obtained by adding all instances of the T-schema to the morphology of the object theory O, according to Tarski, has to be regarded as 'highly incomplete' as it 'would lack the most important and most fruitful general theorems' such as the general principle of contradiction, which states that it is not the case both a sentence and a its negation belong to the class of true sentences of $\mathcal{L}_{\mathsf{O}}$. However, the theory is trivially adequate in the sense of *Convention T*.[10] Therefore is seems that the satisfiability of *Convention T* is only a necessary condition for an adequate theory of truth, as also the provability of the general principle of contradiction is required.

Halbach's concerns can be also be appreciated if we move from theories in the style of TB to theories structured like $\mathsf{TBD[O]}$. The latter is in fact trivially a *materially adequate* theory of truth for the language of O, but again it cannot prove the general principle of contradiction for O. The proof of Observation 1 below derives directly from Tarski's remarks.

To carry out the proof, we also need to define the system $\mathsf{H{\cdot}O}^{sq}$ formulated in the language $\mathcal{L}_{\mathsf{TD}}$. For precise definitions, we refer to the next chapter in which the theory $\mathsf{H{\cdot}O}$, of which $\mathsf{H{\cdot}O}^{sq}$ is a simple extension, will play an important role (cfr. §5.2.1). $\mathsf{H{\cdot}O}^{sq}$ is just $\mathsf{H{\cdot}O}$ formulated in $\mathcal{L}_{\mathsf{TD}}$ plus SQ. For now, it is sufficient to say that $\mathsf{H{\cdot}O}^{sq}$ contains the axioms O, the axioms of HF with HF3 expanded to the entire $\mathcal{L}_{\mathsf{TD}}$ but still applied to variables of syntactic sort s, and the axiom SQ.

Observation 1. $\mathsf{TBD[O]} \nvdash \forall k\big(\mathsf{Sent}^s_{\mathcal{L}_{\mathsf{O}}}(k) \to (\neg\forall a\, \mathsf{Sat}(a,k) \lor \neg\forall a\, \mathsf{Sat}(a, \dot{\neg}k))\big)$.

[10]The fact that *Convention T* is restricted to definitions of truth is also not a problem, as in the theory of truth so-constructed is easily definable a truth predicate if one has already a predicate satisfying the Tarski-equivalences. [Halbach 2011, p. 19-20] considers also less trivial examples of a definition of truth yielding T-sentences but not the general principle of contradiction.

Proof. Let

$$\mathsf{GPC} :\leftrightarrow \forall k\big(\mathsf{Sent}^s_{\mathcal{L}_\mathsf{O}}(k) \to (\neg\forall a\,\mathsf{Sat}(a,k) \vee \neg\forall a\,\mathsf{Sat}(a,\dot\neg k))\big)$$

and assume $\mathsf{TBD[O]} \vdash \mathsf{GPC}$. Let $\mathsf{TBD^*[U]}$ be the finite subsystem of $\mathsf{TBD[O]}$ employed in the proof, where $\mathsf{U}\subseteq\mathsf{O}$: in particular, only finitely many instances of Tbd are employed in the proof applying to finitely many sentences $\sigma_1, ..., \sigma_n$ of $\mathcal{L}_\mathsf{O}$. Let us choose some $\mathcal{L}_\mathsf{O}$-sentence χ such that χ is not among the σ_i's. We thus have

$$\mathsf{TBD^*[U]} \vdash \neg\forall a\,\mathsf{Sat}(a,\ulcorner\chi\urcorner) \vee \neg\forall a\,\mathsf{Sat}(a,\dot\neg\ulcorner\chi\urcorner) \tag{4.3}$$

Now we can interpret $\mathsf{TBD^*[U]}$ in $\mathsf{H{\cdot}O}^{sq}$ in such a way that, according to the interpretation, each occurrence of $\mathsf{Sat}(a,k)$ in the proof can be translated as

$$a = a \wedge (k = \ulcorner\sigma_1\urcorner \wedge \sigma_1) \vee ... \vee (k = \ulcorner\sigma_n\urcorner \wedge \sigma_n) \vee (k = \ulcorner\chi\urcorner) \vee (k = \dot\neg\ulcorner\chi\urcorner), \tag{4.4}$$

where $\ulcorner\sigma_i\urcorner$ denotes the term of $\mathcal{L}_\mathsf{HF}$ coding the $\mathcal{L}_\mathsf{O}$-sentence σ_i. Let us call $\mathcal{T}(a,k)$ the formula (4.4). By (4.3),

$$\mathsf{H{\cdot}O}^{sq} \vdash \neg\mathcal{T}(a,\ulcorner\chi\urcorner) \vee \neg\mathcal{T}(a,\dot\neg\ulcorner\chi\urcorner),$$

but also

$$\mathsf{H{\cdot}O}^{sq} \vdash \neg(\neg\mathcal{T}(a,\ulcorner\chi\urcorner) \vee \neg\mathcal{T}(a,\dot\neg\ulcorner\chi\urcorner)).$$

qed.

Therefore, also in an axiomatic setting that we judge as a better approximation to the Tarskian picture of the metatheory, Tarski's adequacy requirements does not seem to be entirely satisfactory.

4.3 Compositional Truth: the theory $\mathsf{CTD}[\mathsf{O}]$

Definitions 22 and 23 of the *Wahrheitsbegriff* provide us with a recursive definition of the notion of true sentence for the object theory O in the metatheory by means of a relation of satisfaction between a sequence of variable assignments and a metatheoretic name of a formula of $\mathcal{L}_{\mathsf{O}}$. Tarski, however, did not discuss an axiomatisation based upon this inductive definition. He appeared instead quite skeptical about the impact that his definition could have on the study of natural language: the main reason behind his skepticism obviously consisted in semantic paradoxes. According to Tarski, if one wanted to extend his account of the notion of truth for formalised languages to natural language, it would be necessary to decompose the language in a hierarchical structure of sublanguages in which the all-present ambiguities of natural language are reduced to the minimum; this process, however, could distort the very nature of language.[11]

[Davidson 1967] on the contrary seemed to attribute to the 'compositional' clauses turned into axioms a crucial role in the context of his account of meaning of natural language expressions. The impact that a formal system constructed by adding to the object theory compositional axioms for truth can have on the understanding of natural language is not clear yet; furthermore, there is no doubt that Davidson's interest on this axiomatisation called for a precise formulation which was never carried out by Davidson himself or by his disciples. Theories in the style of CT (see Chapter 2) seem to be good candidates to capture Davidson's aims.

Besides historical considerations it seems also that a compositional axiomatisation of the notion of truth addresses an important aspect of our intuitions concerning truth—at least as basic as the disquotational intuition—namely the fact that it nicely interacts with logical constants.[12] Furthermore, the compositional nature of truth, as well as of other semantic notions such as reference and meaning, seems to be a necessary prerequisite

[11] Cfr. [Tarski 1956a, p. 165 and p. 267].
[12] Cfr. [Horsten 2011, p. 70].

for decomposing the structure and understanding the truth-conditions of a declarative statement.[13]

The theory CT[O], whose axioms are displayed in Table 2.1 of Chapter 2, represents one possibility of constructing an axiomatic theory of truth by turning Tarski's clauses into axioms. Another option, definitely more in line with our general purpose, is the theory CTD[O], formulated in $\mathcal{L}_{\mathsf{TD}}$, that will be shortly introduced. CTD[O] is constructed by extending $\mathsf{H{\cdot}O}^{sq}$ by means of compositional axioms for satisfaction. Preliminary to the formulation of CTD[O] are some notational remarks.

The choice of a binary satisfaction predicate instead of a unary truth predicate is justified by the necessity of the presence of a third domain of sequences of objects in the sense of a model of the object theory to connect the realm of syntactic and mathematical objects. Sequences will thus enable us to formulate the truth axioms for atomic formulas and for universally quantified formulas of $\mathcal{L}_{\mathsf{O}}$. To axiomatise a unary truth predicate instead we would require (i) that $\mathcal{L}_{\mathsf{HF}}$ has names for all objects in the domain of discourse of the object theory O and (ii) that we can introduce a denotation function of type $s \to o$ which mimics the role of the evaluation function given in Chapter 2 (§1.2.2) in $I\Sigma_1$. Moreover, if a unary truth predicate is employed, the truth axiom for universal quantification will state that (the code in $\mathcal{L}_{\mathsf{HF}}$) of a universally quantified formula $\forall v_i \varphi$ of $\mathcal{L}_{\mathsf{O}}$ is true if and only if all its substitutional instances are true. We would thus also require (iii) a substitution function of type $(s, s, o) \to s$. However, defining these special functions would lead us to unnecessary complications. An axiomatisation of a binary satisfaction predicate, in this respect, looks much more natural and widely applicable.

An important role in the axiomatisation of CTD[O] will be played by the functions $a(k)$ and $D(a, k)$ of type $(sq, s) \to o$. These functions take a sequence of variable assignments—living in a different domain (see Figure 4.1)—and a finite set and return a mathematical object in the sense of a model of O. The function $D(a, k)$ is dispensable in the case in which

[13] Cfr. [Leitgeb 2007, p. 280].

$\mathcal{L}_{\mathsf{O}}$, like in the case of the language $\mathcal{L}_{\in}$ of set theory, has no terms. In those cases, with R an arbitrary relation in $\mathcal{L}_{\mathsf{O}}$, the truth axiom for atomic formulas will be

$$\text{CTDR} \qquad \forall a \forall i_1, \ldots, i_n \big(\mathsf{Sat}(a, \underset{\cdot}{R} v_1, \ldots, v_n) \leftrightarrow R(a(i_1), \ldots, a(i_n))\big),$$

where $i_1, \ldots, i_n$ range over finite ordinals in the sense of HF. If on the other hand $\mathcal{L}_{\mathsf{O}}$ has terms, then we have to generalise the function $a(k)$ to the function $D(a, k)$, defined for constants and complex terms of $\mathcal{L}_{\mathsf{O}}$ and governed by the axioms

CTDv $\quad \forall a \forall k(D(a, \ulcorner v_k \urcorner) = a(k))$

CTDc $\quad \forall a \forall k(D(a, \underset{\cdot}{c}_k) = c_k)$

CTDf $\quad \forall a \forall \ulcorner \mathrm{t}_1 \urcorner, \ldots, \forall \ulcorner \mathrm{t}_n \urcorner \big(D(a, \underset{\cdot}{f}{}^n_i \mathrm{t}_1, \ldots, \mathrm{t}_n) = f^n_i(D(a, \ulcorner \mathrm{t}_1 \urcorner), \ldots, D(a, \ulcorner \mathrm{t}_n \urcorner))\big).$

The axioms for propositional connectives present no peculiarities. For universally quantified formulas, we just adapt Tarski's clause, given in Definition 22 of the *Wahrheitsbegriff*, to the present setting. CTD$\forall$ will then simply formalise the clause stipulating that a universally quantified formula $\forall v_i \varphi$ is satisfied by a sequence a if and only if all sequences b that differ from a in what they assign to the i^{th} variable satisfy φ; namely

$$\text{CTD}\forall \quad \forall a \forall i \forall \ulcorner \varphi \urcorner (\mathsf{Sat}(a, \underset{\cdot}{\forall} v_i \ulcorner \varphi \urcorner) \leftrightarrow \forall b(\forall j(j \neq i \rightarrow a(j) = b(j)) \rightarrow \mathsf{Sat}(b, \ulcorner \varphi \urcorner))).$$

We can finally present our official definition of CTD[O], standing for compositional truth with 'disentangled' syntax:

Definition 4.3 (CTD[O])**.** The theory CTD[O] is formulated in the three sorted-language $\mathcal{L}_{\mathsf{TD}}$. Its axioms are displayed in Table 4.1.

On occasion, we will consider the subsystem

$$\mathsf{CTD[O]}\!\upharpoonright := \mathsf{CTD[O]} - \mathsf{HF3}^*.$$

(I) Axioms of O;

(II) HF1, HF2;

(III) Axiom governing sequences of variable assignments:

SQ $\forall a \forall x \forall j \exists b(\forall i(i \neq j \to a(i) = b(i)) \wedge b(j) = x)$

(IV) Axioms for denotation and satisfaction:

CTDv $\forall a \forall k(D(a, \underset{\cdot}{v}_k) = a(k))$

CTDc $\forall a \forall k(D(a, \underset{\cdot}{c}_k) = c_k)$

CTDf $\forall a \forall \ulcorner t_1 \urcorner, \ldots, \forall \ulcorner t_n \urcorner \big(D(a, \underset{\cdot}{f}{}^n_i t_1, \ldots, t_n) = f_i^n(D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner))\big)$

for each function symbol f_i^n of $\mathcal{L}_O$.

CTDR $\forall a \forall \ulcorner t_1 \urcorner, \ldots, \forall \ulcorner t_n \urcorner \big(\mathsf{Sat}(a, \underset{\cdot}{R}{}^n_i t_1, \ldots, t_n) \leftrightarrow R_i^n(D(a, \ulcorner t_1 \urcorner), \ldots, D(a, \ulcorner t_n \urcorner))\big)$

for each relation symbol R_i^n in $\mathcal{L}_O$.

CTD¬ $\forall a \forall \ulcorner \varphi \urcorner \big(\mathsf{Sat}(a, \underset{\cdot}{\neg} \ulcorner \varphi \urcorner) \leftrightarrow \neg \mathsf{Sat}(a, \ulcorner \varphi \urcorner)\big)$

CTD → $\forall a \forall \ulcorner \varphi \urcorner \forall \ulcorner \psi \urcorner \big(\mathsf{Sat}(a, \ulcorner \varphi \urcorner \underset{\cdot}{\to} \ulcorner \psi \urcorner) \leftrightarrow \mathsf{Sat}(a, \ulcorner \varphi \urcorner) \to \mathsf{Sat}(a, \ulcorner \psi \urcorner)\big)$

CTD∀ $\forall a \forall \ulcorner \varphi \urcorner \forall i \big(\mathsf{Sat}(a, \underset{\cdot}{\forall} v_i \ulcorner \varphi \urcorner) \leftrightarrow \forall b(\forall j(j \neq i \to a(j) = b(j)) \to \mathsf{Sat}(b, \ulcorner \varphi \urcorner))\big)$

(V) HF3* $A(0) \wedge \forall k \forall l(A(k) \wedge A(l) \to A(k \lhd l)) \to \forall k A(k)$

where k is a variable of sort s and A is a formula of $\mathcal{L}_{\mathsf{TD}}$

Table 4.1: Axioms of CTD[O].

Axiom HF3* deserves some additional explanation: in Chapter 2 we noticed that in the case of schematically axiomatised object theories it is possible to distinguish, in informal metamathematical discussion, between a *syntactic* and a *mathematical* role of the schemata of the theory of truth constructed in the usual way. Since the theory of syntax is identified with the object theory, however, two instances of the same schema playing different roles are indistinguishable from the point of view of the theory of truth, even though they differ in our informal metamathematical reflection. Moreover, we observed that the provability of the so-called Global Reflection Principle for a schematically axiomatised object theory O, that is the full, explicit assertion of the soundness of O,

$$\mathsf{GRP}^o_{\mathsf{O}} :\leftrightarrow \forall x(\mathsf{Sent}^o_{\mathcal{L}_{\mathsf{O}}} \wedge \mathsf{Bew}^o_{O}(x) \to Tx) \tag{4.5}$$

essentially relies on the fact that the syntactic version of the schemata of the object theory extended with to the truth predicate are employed as if they were mathematical.

The schema HF3*, being restricted to variables ranging over hereditarily finite sets, is devised to preserve the possibility of performing syntactic arguments involving the notion of truth, such as inductions on the length of the proofs of O or on the complexity of expressions of $\mathcal{L}_{\mathsf{O}}$ but, at the same time, to eliminate the possibility of formulating arguments which essentially rely on a mathematical role of the schemata, such as arguments aiming at the truth of a formula of $\mathcal{L}_{\mathsf{O}}$ at each substitutional instance. To this purpose the following lemma, which is an immediate consequence of Proposition 3.2, will be quite useful:

Lemma 4.1 (Complete Ordinal Induction)**.** For any formula $A(n)$ of $\mathcal{L}_{\mathsf{TD}}$ with $\mathsf{Ord}(n)$ (we recall that n is of sort s):

$$\mathsf{CTD[O]} \vdash \forall n((\forall m < n)\, A(m) \to A(n)) \to \forall n\, A(n).$$

4.4 'Syntactic' Soundness Claims and Consistency

In the present section some noticeable theorems *of* CTD[O] and metatheorems *on* CTD[O] will be introduced. We will first prove that UTDB[O], and thus TBD[O], are subsystems of CTD[O]. We then prove some useful syntactic facts about O, which exemplify the role of the syntactic principle HF3*. We finally consider an interesting 'splitting' phenomenon involving consistency statements for O: in particular, in CTD[O] we can construct consistency statements for O formalised in the language of O, but also consistency statements formalised in the language $\mathcal{L}_{\mathsf{HF}}$ of our 'detached' syntax theory. In the present section, we will be mainly concerned with 'syntactic' consistency statements.

Tarski biconditionals. Let $\ulcorner\varphi\urcorner[\ulcorner \mathrm{t}_1\urcorner/1, \dots, \ulcorner \mathrm{t}_n\urcorner/n]$ denote the obvious generalisation of the formal substitution function defined by (3.22) and let $b \overset{i}{\sim} a$ denote that the sequence b differs from a in what it assigns to its i^{th} element at most. The next lemma shows that UTDB[O] and TDB[O] are subtheories of CTD[O].

Lemma 4.2 (Uniform Disquotation.)**.** For all formulas $\varphi(v_1, \dots, v_n)$ of $\mathcal{L}_{\mathsf{O}}$ with only the variables displayed free, CTD[O]$\restriction$ proves

$$\forall a \forall \ulcorner \mathrm{t}_1\urcorner, \dots, \ulcorner \mathrm{t}_n\urcorner(\mathsf{Sat}(a, \ulcorner\varphi\urcorner[\ulcorner \mathrm{t}_1\urcorner/1, \dots, \ulcorner \mathrm{t}_n\urcorner/n]) \leftrightarrow \varphi(D(a, \ulcorner \mathrm{t}_1\urcorner), \dots, D(a, \ulcorner \mathrm{t}_n\urcorner)))$$

Sketch Proof. The proof is by metatheoretic induction on the complexity of $\varphi(v_1, \dots, v_n)$. The crucial case is the one in which $\varphi(v_1, \dots, v_n)$ is $\forall v_j \psi(v_j, v_1, \dots, v_n)$.

For the left-to-right direction, $\mathsf{Sat}(a, \underset{\cdot}{\forall} v_j \ulcorner\psi\urcorner[\ulcorner \mathrm{t}_1\urcorner/1, \dots, \ulcorner \mathrm{t}_n\urcorner/n])$ entails, by CTD$\forall$, that for all $b \overset{j}{\sim} a$, $\mathsf{Sat}(b, \ulcorner\psi(\mathrm{t}_1, \dots, \mathrm{t}_n)\urcorner)$. By the induction hypothesis,

$$\psi(b(j), D(b, \ulcorner \mathrm{t}_1\urcorner), \dots, D(b, \ulcorner \mathrm{t}_n\urcorner)).$$

Since no v_j can appear in the t_i's, by SQ we have $\psi(x, D(a, \ulcorner \mathrm{t}_1\urcorner), \dots, D(a, \ulcorner \mathrm{t}_n\urcorner))$, with x an arbitrary term ‘of $\mathcal{L}_O$ not occurring in the t_i; thus

$$\forall v_j \psi(v_j, D(a, \ulcorner \mathrm{t}_1 \urcorner), ..., D(a, \ulcorner \mathrm{t}_n \urcorner)).$$

For the converse, we need to prove $\mathsf{Sat}(b, \ulcorner \psi(\mathrm{t}_1, ..., \mathrm{t}_n) \urcorner)$ for any $b \overset{j}{\sim} a$. We consider an arbitrary $b \overset{j}{\sim} a$: by assumption we have $\forall v_j \psi(v_j, D(b, \ulcorner \mathrm{t}_1 \urcorner), ..., D(b, \ulcorner \mathrm{t}_n \urcorner))$ which becomes, by SQ,

$$\psi(b(j), D(b, \ulcorner \mathrm{t}_1 \urcorner), ..., D(b, \ulcorner \mathrm{t}_n \urcorner)). \tag{4.6}$$

But by induction hypothesis (4.6) is just $\mathsf{Sat}(b, \ulcorner \psi(\mathrm{t}_1, ..., \mathrm{t}_n) \urcorner)$ for an arbitrary $b \overset{j}{\sim} a$, that is

$$\mathsf{Sat}(a, \underset{\cdot}{\forall} v_j \ulcorner \psi \urcorner [\ulcorner \mathrm{t}_1 \urcorner / 1, ..., \ulcorner \mathrm{t}_n \urcorner / n]).$$

qed.

Syntactic lemmata. Given the syntactic principle HF3*, CTD[O] can prove nontrivial facts concerning O which, in the usual setting, require the full axiomatisation of Tarskian truth, that is the theory that we labeled CT[O]. These minor results show once more how, in the customary setting, it is impossible to isolate statements of a purely syntactic nature without appealing to informal metatheoretic considerations.

We first prove that substitution of terms with identical values—in the sense of our ‘bridging’ function $D(a, k)$—preserves the truth of the formulas in which they occur. By recursion on the rank of hereditarily finite sets (Cfr. Chapter 2, Lemma 3.5), we define the function $\mathsf{lc}^s(k)$ that measures the complexity of the formula coded by k (we recall from Chapter 2 that $\overline{n}$ denotes the predecessor of the ordinal n):

$$\mathsf{lc}^s(k) = \begin{cases} 0, & \text{if } \mathsf{AtFml}^s_{\mathcal{L}_O}(k) \text{ or } \neg\mathsf{Fml}^s_{\mathcal{L}_O}(k); \\ \\ n, & \text{if } (k = l \dot{\to} m \wedge \mathsf{max}^s(\mathsf{lc}^s(m), \mathsf{lc}^s(l)) = \overline{n}) \vee \\ & (k = \dot{\neg} l \wedge \mathsf{lc}^s(l) = \overline{n}) \vee (k = \dot{\forall} v_i l \wedge \mathsf{lc}^s(l) = \overline{n}) \\ & \text{for formulas } l, m \text{ and ordinals } i. \end{cases}$$

Proposition 4.1 (Regularity)**.**

$$\begin{aligned} \mathsf{CTD[O]} \vdash \forall a \forall \ulcorner \varphi(v_i) \urcorner\ \forall \ulcorner \mathsf{t}_1 \urcorner, \ulcorner \mathsf{t}_2 \urcorner \big(D(a, \ulcorner \mathsf{t}_1 \urcorner) = D(a, \ulcorner \mathsf{t}_2 \urcorner) \to \\ (\mathsf{Sat}(a, \ulcorner \varphi \urcorner [\ulcorner \mathsf{t}_1 \urcorner / i]) \leftrightarrow \mathsf{Sat}(a, \ulcorner \varphi \urcorner [\ulcorner \mathsf{t}_2 \urcorner / i])) \big) \end{aligned}$$

Proof. If k codes an atomic formula, thus $\mathsf{lc}^s(k) = 0$, then for any $\ulcorner \mathsf{t}_1 \urcorner$, $\ulcorner \mathsf{t}_2 \urcorner$ and a and ordinal i,

$$D(a, \ulcorner \mathsf{t}_1 \urcorner) = D(a, \ulcorner \mathsf{t}_2 \urcorner) \to \mathsf{Sat}(a, k[\ulcorner \mathsf{t}_1 \urcorner / i]) \leftrightarrow \mathsf{Sat}(a, k[\ulcorner \mathsf{t}_2 \urcorner / i]). \tag{4.7}$$

To obtain (4.7) we first notice that, for any term t of $\mathcal{L}_O$, and for all a and some constant c in $\mathsf{Const}_{\mathcal{L}_O}$,

$$\mathsf{CTerm}^s_{\mathcal{L}_O}(\ulcorner \mathsf{t}(c) \urcorner) \wedge D(a, \ulcorner \mathsf{t}_1 \urcorner) = D(a, \ulcorner \mathsf{t}_2 \urcorner) \to D(a, \ulcorner \mathsf{t}(\mathsf{t}_1) \urcorner) = D(a, \ulcorner \mathsf{t}(\mathsf{t}_2) \urcorner). \tag{4.8}$$

Then (4.7) follows from $\mathsf{CTD}R$.

The proof is then completed by applying HF3* to the ordinal n in the formula

$$\begin{aligned} \forall a \forall \ulcorner \varphi(v_i) \urcorner \big(\mathsf{lc}^s(\ulcorner \varphi(0) \urcorner) \leq n \to (D(a, \ulcorner \mathsf{t}_1 \urcorner) = D(a, \ulcorner \mathsf{t}_2 \urcorner) \to \\ (\mathsf{Sat}(a, \ulcorner \varphi \urcorner [\ulcorner \mathsf{t}_1 \urcorner / i]) \leftrightarrow \mathsf{Sat}(b, \ulcorner \varphi \urcorner [\ulcorner \mathsf{t}_2 \urcorner / i]))) \big) \end{aligned}$$

The base case is covered by (4.7). For the induction step, the compositional axioms and the induction hypotheses are sufficient to cover all cases. For instance, if $\varphi(v_i)$ is $\neg\psi(v_i)$, we have that

$$\begin{aligned}
\mathsf{Sat}(a, \underset{\cdot}{\neg}\ulcorner\psi\urcorner[\ulcorner \mathsf{t}_1\urcorner/i]) &\leftrightarrow \neg\mathsf{Sat}(a, \ulcorner\psi\urcorner[\ulcorner \mathsf{t}_1\urcorner/i]) && \mathsf{CTD}\neg\\
&\leftrightarrow \neg\mathsf{Sat}(a, \ulcorner\psi\urcorner[\ulcorner \mathsf{t}_2\urcorner/i]) && \text{by induction hypothesis}\\
&\leftrightarrow \mathsf{Sat}(a, \underset{\cdot}{\neg}\ulcorner\psi\urcorner[\ulcorner \mathsf{t}_2\urcorner/i]) && \mathsf{CTD}\neg
\end{aligned}$$

qed.

Let now $\vec{v}$ denote a sequence s of length n such that for all $m < n$, $s_m = (v)_m$; that is s_m is the m^{th} element in the sequence $\vec{v}$ whose elements are variables. Therefore we have

$$\underset{\cdot}{\forall}\vec{v}\ \ulcorner\varphi\urcorner = (\ulcorner\forall\urcorner, (\vec{v})_1, (\ulcorner\forall\urcorner, (\vec{v})_2, (...(\ulcorner\forall\urcorner, (\vec{v})_m, \ulcorner\varphi\urcorner)...))),$$

where $\vec{v}$ is a sequence of ordinals of length $m \lhd m$. The next lemma is an immediate consequence of the definitions. It states that $\underset{\cdot}{\forall} v_i k$ is a sentence if and only if k codes a $\mathcal{L}_\mathrm{O}$-formula with free variables among elements of $\vec{v}$.

Lemma 4.3. Let $\vec{v}$ be a sequence. Then $\mathsf{CTD}[\mathrm{O}]$ proves:

$$\begin{aligned}
&\Big(\mathsf{Fml}^s_{\mathcal{L}_\mathrm{O}}(k) \wedge \mathsf{Seq}^s(\vec{v}, n) \wedge (\forall i)\ [\mathsf{var}^s_{\mathcal{L}_\mathrm{O}}(i) \wedge \mathsf{FreeF}^s(i,k) \to \exists j < n(i = (\vec{v})_j)\Big)\\
&\leftrightarrow \mathsf{Sent}^s_{\mathcal{L}_\mathrm{O}}(\underset{\cdot}{\forall}\vec{v}k).
\end{aligned}$$

We can also prove a generalisation of $\mathsf{CTD}\forall$. Let the expression $b \overset{1,\dots,n}{\sim} a$ denote that the sequence b differs from a in what it assigns to the $\mathcal{L}_\mathrm{O}$-variables $v_1, \dots, v_n$; we also recall that

$$\forall\ulcorner\varphi(\vec{v})\urcorner(...) :\leftrightarrow \forall k(\mathsf{Sent}^s_{\mathcal{L}_\mathrm{O}}(\underset{\cdot}{\forall}\vec{v}k) \to ...)$$

Proposition 4.2. Let $\vec{v}$ be a sequence of length $n \lhd n$, then

$$\mathsf{CTD[O]} \vdash \forall \vec{v} \forall a \forall \ulcorner \varphi(\vec{v}) \urcorner (\mathsf{Sat}(a, \underset{.}{\forall} \vec{v} \ulcorner \varphi \urcorner) \leftrightarrow \forall b \overset{(\vec{v})_1, \ldots, (\vec{v})_n}{\sim} a(\mathsf{Sat}(b, \ulcorner \varphi \urcorner))).$$

Proof. The result follows by applying HF3*—in particular, the extended schema of complete induction (Cfr. Lemma 4.1)—to the ordinal n in the formula

$$\forall a \forall \ulcorner \varphi(\vec{v}) \urcorner \forall i \leq n \big(\mathsf{FreeF}((\vec{v})_i, \ulcorner \varphi \urcorner) \rightarrow (\mathsf{Sat}(a, \underset{.}{\forall} \vec{v} \ulcorner \varphi \urcorner) \leftrightarrow \\ \forall b \overset{(\vec{v})_1, \ldots, (\vec{v})_n}{\sim} a(\mathsf{Sat}(b, \ulcorner \varphi \urcorner)))\big).$$

The result then follows by $\mathsf{CTD}\forall$. *qed.*

Soundness Claims and Consistency Statement(s): finitely axiomatised theories. In §1.2.2 we employed an induction on the length of the derivations in O to prove the Global Reflection Principle for O—actually for PA and ZF—that is the assertion that all theorems of O are true

$$\mathsf{GRP}^o_{\mathsf{O}} \qquad \forall \overline{\sigma}(\mathsf{Sent}^o_{\mathcal{L}_\mathsf{O}}(\overline{\sigma}) \wedge \mathsf{Bew}^o_{\mathsf{O}}(\overline{\sigma}) \rightarrow T\overline{\sigma})$$

belonging to the language of $\mathcal{L}_\mathsf{O}$.[14] The argument can now be naturally transferred in CTD[O], by virtue of HF3*, if O is finitely axiomatised. The reasoning does not exhibit any substantial difference with respect to the usual setting: the dissimilarities regard only the language in which the global reflection principle is formulated. Let us define

$$\mathsf{GRP}^s_{\mathsf{O}} :\leftrightarrow \forall a \forall \ulcorner \varphi \urcorner (\mathsf{Bew}^s_{\mathsf{O}}(\ulcorner \varphi \urcorner) \rightarrow \mathsf{Sat}(a, \ulcorner \varphi \urcorner)).$$

The following proposition does not depend on the signature of $\mathcal{L}_\mathsf{O}$.

[14] Recall our notational convention: the line above an expression denotes the term coding it within $\mathcal{L}_\mathsf{O}$.

Proposition 4.3. Let O be finitely axiomatised, then $\mathsf{CTD}[\mathsf{O}] \vdash \mathsf{GRP}^s_{\mathsf{O}}$.

Proof. By complete ordinal induction on the n in the formula (we recall from Chapter 2 that $\mathsf{B}^s_{\mathsf{O}}(k, n, \ulcorner\varphi\urcorner)$ expresses that k is a proof of $\ulcorner\varphi\urcorner$ of length n):

$$\forall a \forall \ulcorner\varphi\urcorner (\exists k(\mathsf{B}^s_{\mathsf{O}}(k, n, \ulcorner\varphi\urcorner)) \rightarrow \mathsf{Sat}(a, \ulcorner\varphi\urcorner)).$$

for all $\ulcorner\varphi\urcorner$. When φ is an axiom, the claim is immediately obtained by Lemma 4.2. If φ is obtained by MP from ψ and $\psi \rightarrow \varphi$, the claim follows by induction hypothesis and $\mathsf{CTD}\rightarrow$. If φ is $\forall v_i \psi$ and it is obtained by Gen from ψ, where v_i does not occur free in ψ, then the claim follows immediately by induction hypothesis, as if ψ is satisfied by an arbitrary sequence a and v_i is not free in ψ, then ψ is also satisfied by all $b \overset{i}{\sim} a$. *qed.*

Let now $\mathsf{Con}^s_{\mathsf{O}} :\leftrightarrow \neg\mathsf{Bew}^s(\ulcorner\xi\urcorner)$, where ξ is a refutable $\mathcal{L}_{\mathsf{O}}$-sentence in O. By instantiating ξ in $\mathsf{GRP}^s_{\mathsf{O}}$, we obtain $\mathsf{Bew}^s_{\mathsf{O}}(\ulcorner\xi\urcorner) \rightarrow \mathsf{Sat}(a, \ulcorner\xi\urcorner)$, thus $\mathsf{Bew}^s_{\mathsf{O}}(\ulcorner\xi\urcorner) \rightarrow \xi$ by Lemma 4.2. Therefore, since ξ is refutable in $\mathsf{CTD}[\mathsf{O}]$, we have

Corollary 4.1. $\mathsf{CTD}[\mathsf{O}] \vdash \mathsf{Con}^s_{\mathsf{O}}$.

Lemma 4.3 and Corollary 4.1 tell us that the $\mathsf{CTD}[\mathsf{O}]$, when O is finitely axiomatised, behaves along the lines of the full compositional theory $\mathsf{CT}[\mathsf{O}]$.

If we now move to the case in which O is a *schematically axiomatised* theory, in the sense of section 2.3.1, we will see that $\mathsf{CTD}[\mathsf{O}]$ behaves in a rather different fashion, if compared to the usual setting. The proofs of the metatheorems that will be presented in the remaining part of this section, however, cannot reach the generality displayed by the proofs of the theorems of $\mathsf{CTD}[\mathsf{O}]$ considered so far. We will thus concentrate on two well-known schematically axiomatisable theories formulated in different languages, Peano Arithmetic and Zermelo-Fraenkel set theory. It may be conjectured that the techniques employed, with suitable but minor modifications, might be applied to different theories.

Interpretations. Before tackling the metatheorems just announced, we introduce some useful notions concerning relative interpretations.[15]

Let S be a theory in $\mathcal{L}_{\mathsf{S}}$ and T a theory in $\mathcal{L}_{\mathsf{T}}$. A *relative translation* $\tau\colon \mathcal{L}_{\mathsf{S}} \to \mathcal{L}_{\mathsf{T}}$ is understood as a pair (δ, F) where δ is a formula of $\mathcal{L}_{\mathsf{T}}$ with one free variable and F is a mapping that associates to each relation symbol R of $\mathcal{L}_{\mathsf{S}}$ of arity n a formula $F(R)(v_1, ..., v_n)$ of $\mathcal{L}_{\mathsf{T}}$; to each function symbol f of $\mathcal{L}_{\mathsf{S}}$ of arity n a formula $F(f)(v_1, ..., v_{n+1})$ of $\mathcal{L}_{\mathsf{T}}$ and to each constant c of $\mathcal{L}_{\mathsf{S}}$ a formula $F(c)(v_1)$. The mapping F induces the extended mapping τ which associates to every formula $\varphi(v_1, ..., v_n)$ of $\mathcal{L}_{\mathsf{S}}$ its translation $(\varphi)^\tau(v_1, ..., v_n)$ in the following way:

$$
\begin{aligned}
&(c = v_i)^\tau :\leftrightarrow F(c)(v_i)\\
&(f(v_1, ..., v_n) = y)^\tau :\leftrightarrow F(f)(v_1, ..., v_n, y)\\
&(R(v_1, ..., v_n, y))^\tau :\leftrightarrow F(R)(v_1, ..., v_n, y)\\
&(\neg\varphi)^\tau :\leftrightarrow \neg(\varphi)^\tau\\
&(\varphi \to \psi)^\tau :\leftrightarrow (\varphi)^\tau \to (\psi)^\tau\\
&(\forall v_i \varphi)^\tau :\leftrightarrow \forall v_i(\delta(v_i) \to (\varphi)^\tau).
\end{aligned}
$$

A *relative interpretation* of S in T is specified by a triple $(\mathsf{S}, \tau, \mathsf{T})$, where $\tau = (\delta, F)$ is a relative translation of $\mathcal{L}_{\mathsf{S}}$ in $\mathcal{L}_{\mathsf{T}}$. In particular, we require that $\mathsf{T} \vdash \exists x\, \delta(x)$, that is that the domain of the quantifiers is nonempty; that for any n-ary function symbol f of $\mathcal{L}_{\mathsf{S}}$,

$$\mathsf{T} \vdash \forall v_1, ..., v_n(\delta(v_1) \wedge ... \wedge \delta(v_n) \to \exists! v_j(\delta(v_j) \wedge F(f)(v_1, ..., v_n, v_j)))$$

with $F(f)$ as above and consequently that T proves $\exists! v_i\ (\delta(v_i) \wedge F(c)(v_i)$ for each constant c of $\mathcal{L}_{\mathsf{S}}$. Finally, if σ is an axiom of S,[16] we require that $\mathsf{T} \vdash (\sigma)^\tau$. By a straighforward

[15] We follow [Hájek&Pudlák 1993] and [Visser 2011] to some extent.
[16] Here we assume that axioms of S are sentences.

induction on the length of the derivation in S, it follows that for all sentences σ provable in S, $\mathsf{T} \vdash (\sigma)^\tau$. We write $\mathsf{S} \leqslant \mathsf{T}$ for the interpretability relation, that is the relation that holds among S and T if there is an interpretation of S in T. We also notice that if $\mathsf{S} \leqslant \mathsf{T}$, then we can always construct in any model $\mathcal{M}$ of T an internal model $\mathcal{N}$ of S.

An interpretation is *direct* if identity is mapped to identity and there is no relativisation of quantifiers—for instance, the formula $\delta(x)$ in the definition of the translation can be taken to be just $x = x$. Moreover, we say that an interpretation is *faithful* if and only if for all sentences σ of $\mathcal{L}_\mathsf{S}$, $\mathsf{S} \vdash \sigma$ *if and only if* $\mathsf{T} \vdash (\sigma)^\tau$. We also say that S is *locally interpretable* in T if and only if every finite subsystem of S is interpretable in T. In what follows we will be mainly concerned with the case in which the interpreted theory is formulated in a many-sorted language, whereas the language of the interpreting theory has only one sort. To deal with these cases, in the definition of relative translation we specify possibly different formulas of the language of the interpreting theory relativising quantifiers of the interpreted theory.

Furthermore, when dealing with interpreting theories that are extensions of PA in the language $\mathcal{L}_\mathsf{A}$ of arithmetic, local interpretability is indeed a powerful tool. By the essential reflexivity of PA, in fact, we have

Fact 4.1 (Orey)**.** If a (recursively axiomatisable) system S is locally interpretable in T and T is an extension of PA in the language of PA, then S is interpretable in T.[17]

Soundness Claims and Consistency Statement(s): Zermelo Fraenkel Set Theory. Our aim is to prove that, unlike $\mathsf{CT[ZF]}$, $\mathsf{CTD[ZF]}$ cannot prove the global reflection principle for ZF. For the intended axiomatisation and other relevant properties of ZF, we refer to [Jech 2003].

In ZF, we standardly define the cumulative hierarchy $\mathbb{V}$ by transfinite induction as the union of all V_α for ordinals α.

[17] For a proof of this result, see [Hájek&Pudlák 1993, §III.2.(e)].

Lemma 4.4. Let O be ZF. Then $\mathsf{CTD}[\mathsf{O}]$ is locally interpretable in O.

Proof. Let $\mathsf{CTD}^*[\mathsf{U}]$ be an arbitrary finite subsystem of $\mathsf{CTD}[\mathsf{ZF}]$ in which only finitely many occurrences of $\mathsf{HF3}^*$ and of separation and replacement of ZF occur—where U is the finite subsystem of ZF featuring those schemata. We define a translation τ_1 of $\mathcal{L}_{\mathsf{TD}}$ into $\mathcal{L}_{\in}$.

τ_1 relativises 'mathematical' quantifiers of $\mathcal{L}_{\mathsf{TD}}$, ranging over the domain of discourse of $\mathcal{L}_{\mathsf{O}}$ by means of the formula $x \in V_\alpha$, where we can safely assume that V_α models the axiom of infinity and that $V_\alpha \vDash \mathsf{U}$ is provable in ZF; 'syntactic' quantifiers to the formula $x \in V_\omega$; and quantifiers applying to variables of sort *sq* the formula $x \in V_\alpha^\omega$. Satisfaction in $\mathsf{CTD}^*[\mathsf{U}]$ is translated as satisfaction in V_α according to a sequence living in V_α^ω, expressing that x belongs to the set of functions in V_α with finite ordinal domain. The function $a(i)$ of $\mathcal{L}_{\mathsf{TD}}$ is translated as a projection function for sequences in V_α. Since it will be useful in what follows, we specify the details of the translation, which we will call, abusing of notation, *, of $\mathcal{L}_{\mathsf{HF}}$ into $\mathcal{L}_{\in}$ (see Table 4.2).

$$
\begin{aligned}
&(x = y)^* :\leftrightarrow x = y \\
&(x = 0)^* :\leftrightarrow (\forall y \in x)(y \in y) \\
&(x \lhd y = z)^* :\leftrightarrow (\forall u \in z)(u \in x \vee u = y) \wedge (x \subseteq z) \wedge (y \in z) \\
&(\neg\varphi)^* :\leftrightarrow \neg(\varphi)^* \\
&(\varphi \wedge \psi)^* :\leftrightarrow (\varphi)^* \wedge (\psi)^* \\
&(\forall x \varphi)^* :\leftrightarrow \forall x \in \omega(\varphi)^*.
\end{aligned}
$$

Table 4.2: The 'subinterpretation' *.

We need to verify that ZF proves the translations of the axioms of $\mathsf{CTD}^*[\mathsf{U}]$: axioms of U are obviously satisfied by ZF, and so are $\mathsf{HF1}$ and $\mathsf{HF2}$ and $\mathsf{HF3}^*$. Translations of the axioms for denotation and satisfaction are also satisfied by ZF, by noticing that if a formula φ defines in $\mathcal{L}_{\mathsf{HF}}$ a syntactic relation concerning U, Table 4.2 gives us the means to characterise the syntax of U in ZF in such a way that $(\varphi)^{\tau_1}$ defines the same syntactic relation concerning U in $\mathcal{L}_{\mathsf{U}}$. Logical axioms, including identity axioms, are satisfied by the same token. *qed.*

Lemma 4.4 thus determines a structure $\mathcal{D}$ which, provably in ZF, models the arbitrary finite subsystem $\mathsf{CTD}^*[\mathsf{U}]$ of $\mathsf{CTD}[\mathsf{ZF}]$.

As anticipated, we shall now see that τ_1 gives us the means to highlight a first dissimilarity between the usual setting and the proposed axiomatisation of truth: in particular, Fact 2.3 states that under a usual axiomatisation of compositional truth over ZF,[18] $\mathsf{CT}[\mathsf{ZF}]$ proves $\mathsf{GRP}^{o}_{\mathsf{ZF}}$, that is the consistency statement for ZF formulated in the language of ZF. In $\mathsf{CTD}[\mathsf{ZF}]$ this is not possible.

We first recall a well-know result about the cumulative hierarchy that will be employed below: in ZF every statement concerning $\mathbb{V}$ can be taken to hold of some level in the cumulative hierarchy. We denote with φ^C the relativisation of the $\mathcal{L}_\in$-formula φ to the class C.[19]

Fact 4.2 (Set Theoretic Reflection Principle)**.** Let $\varphi(v_1, \dots, v_n)$ be a $\mathcal{L}_\in$-sentence with the only free variables displayed. Then, with β a limit ordinal,

$$\mathsf{ZF} \vdash \forall\alpha(\exists\beta > \alpha)(\forall v_1, \dots, v_n \in V_\beta)(\varphi(v_1, \dots, v_n) \leftrightarrow \varphi^{V_\beta}(v_1, \dots, v_n)).$$

Theorem 4.1. $\mathsf{CTD}[\mathsf{ZF}] \nvdash \mathsf{Con}^{s}_{\mathsf{ZF}}$.

Proof. Assume $\mathsf{CTD}[\mathsf{ZF}] \vdash \mathsf{Con}^{s}_{\mathsf{ZF}}$. By compactness, consider a finite subsystem $\mathsf{CTD}^*[\mathsf{U}]$ of $\mathsf{CTD}[\mathsf{ZF}]$ that proves $\mathsf{Con}^{s}_{\mathsf{ZF}}$. Construct in ZF a model $\mathcal{D}$ of $\mathsf{CTD}^*[\mathsf{U}]$ by means of the interpretation defined in Lemma 4.4 so that the relativisation of $\mathsf{Con}^{s}_{\mathsf{ZF}}$ to the model of the subsystem S of HF formalising the syntax of U is provable in ZF. Calling the induced interpretation τ_1, we have $\mathsf{ZF} \vdash (\mathsf{Con}^{s}_{\mathsf{ZF}})^{\tau_1}$. Since $\mathsf{Con}^{s}_{\mathsf{ZF}}$ is a $\mathcal{L}_{\mathsf{HF}}$-sentence, quantifiers

[18] Such as the one that can be found in [Fujimoto 2012].
[19] Cfr. [Jech 2003, Ch. I.12] and [Devlin 1984, Ch I.8].

of $(\mathsf{Con}^s_{\mathsf{ZF}})^{\tau_1}$, according to $\mathcal{D}$, are relativised to V_ω. Now the relation $V_\omega \vDash \mathsf{Con}^s_{\mathsf{ZF}}$ is Δ_1^{ZF}, and thus absolute across transitive models. Thus, provided that $V_\omega \subset V_\alpha$, the definition of τ_1 (cfr. Table 4.2) ensures that $(\mathsf{Con}^s_{\mathsf{ZF}})^{\tau_1}$ holds in V_α. Therefore, by Fact 4.2, we can consider $(\mathsf{Con}^s_{\mathsf{ZF}})^{\tau_1}$ a canonical consistency statement for ZF. This contradicts Gödel's Second Incompleteness Theorem.

qed.

Thus, as an immediate consequence,

Corollary 4.2. $\mathsf{CTD}[\mathsf{ZF}] \nvdash \mathsf{GRP}^s_{\mathsf{ZF}}$.

Soundness Claims and Consistency Statement(s): Peano Arithmetic. We now consider the case in which our object theory O is Peano Arithmetic. We shall need to employ the following notions and results.

Let S be a finite set of sentences, and $C = \{c_n : n \in \omega\}$ a countable set of new constants for the language $\mathcal{L}_\mathsf{S}$. We are only interested in the case in which $\mathcal{L}_\mathsf{S}$ is the language of arithmetic. Let $\chi(x, y)$ be the formula of the language $\mathcal{L}_\mathsf{S}$ formalising the inductive construction of the Henkin completion for S;[20] $\chi(x, y)$ thus expresses that there is a sequence z capturing the completion of the set S such that x is an $\mathcal{L}_\mathsf{S} \cup C$-sentence and it occurs at the y^{th} entry in z. We thus define

$$\mathsf{Tr}_\mathsf{S}(x) :\leftrightarrow \exists y\, \chi(x, y).$$

Thus $\mathsf{Tr}_\mathsf{S}(x)$ expresses that the sentence x belongs to the Henkin completion of S. We can finally introduce a sentence Mod_S, expressing that the set defined by $\mathsf{Tr}_\mathsf{S}(x)$ is complete, consistent and that if the code of $\exists v\, \varphi(v)$ belongs to this set, there is a constant c in C such that (the code of) $\varphi(c)$ belongs to the set.[21] In $\mathsf{PA} + \mathsf{Mod}_\mathsf{S}$ it is immediate to verify that $\mathsf{Tr}_\mathsf{S}(x)$ satisfies the Tarskian compositional clauses.

[20] Cfr. [Lindström 2003, p. 79].
[21] *Ibidem.*

Fact 4.3 (Arithmetised Completeness Theorem)**.** Let S a recursive set of sentences and $\mathsf{Con_S}$ a canonical consistency statement for S. Then

$$\mathsf{PA} + \mathsf{Con_S} \vdash \mathsf{Mod_S}.$$

Equivalently, we can formulate Fact 4.3 by claiming that S is interpretable in $\mathsf{PA}{+}\mathsf{Con_S}$.

Lemma 4.5. Let U be an (arbitrary) finite subsystem of PA. $\mathsf{CTD[PA]}$ is locally interpretable in $\mathsf{PA}{+}\mathsf{Mod_U}$.

Proof. As before, take the finite subsystem $\mathsf{CTD^*[U]}$ of $\mathsf{CTD[PA]}$. Moreover, we consider a Δ_1-binumeration of U in PA, so that $\mathsf{Mod_U}$ and the formula $\mathsf{Tr_U}(x)$ can be fixed. Let $c(i)$ represent in PA a primitive recursive function which applies to the i^{th} variable of $\mathcal{L}_\mathsf{U}$ and yields the i^{th} element of C.[22] If $\mathsf{Sent}_{\mathcal{L}_O}(\underset{\cdot}{\forall}\vec{v}\ \overline{\varphi})$, we define in PA a (functional) predicate $\mathsf{cFml}_{\mathcal{L}_\mathsf{U}}(\overline{\varphi}, \vec{v}, y)$ expressing that y codes the result of formally substituting each occurrence of the variables coded in the sequence $\vec{v}$ in φ with $c(\dot{(\vec{v})}_i)$, where $\dot{(\vec{v})}_i$ denotes the numeral for the variable $(\vec{v})_i$. We denote with $\overline{\varphi}[\vec{v}]^\bullet$ the code of this unique sentence. Similarly, we denote with $\overline{t}[\vec{v}]^\circ$ the result of formally replacing the free variables of the $\mathcal{L}_\mathsf{U}$-term t, whose codes appear as values in the sequence $\vec{v}$, with $c(\dot{(\vec{v})}_i)$.

We define the relative translation $\tau_2 = (\delta^o, \delta^{sq}, F)$ of $\mathcal{L}_\mathsf{TD}$ into $\mathcal{L}_\mathsf{PA}$, whose essential components are displayed in Table 4.3.

Since τ_2 maps a many-sorted signature into a one-sorted one, as in Lemma 4.4 we specify more than just one domain for the translation. Let

$$\delta^o(x) :\leftrightarrow \exists y(x = c(y)).$$

$\delta^o(x)$ characterises the domain of our arithmetised model as depicted by $\mathsf{Mod_U}$, and it will

[22] If $\mathcal{L}_\mathsf{U}$ contains constants already, we may partition the set of constants of $\mathcal{L}_{\mathsf{U}\cup\mathsf{C}}$ by taking new constants to have even indices and old ones odd indices.

$$F(0^s) :\leftrightarrow x = 0$$
$$F(\in^s) :\leftrightarrow \mathsf{bit}(x,y) = 1 \qquad \text{Cfr. §3.2.1;}$$
$$F(\lhd) :\leftrightarrow \mathsf{bit}(x,y) = 1 \vee x = z$$
$$F(.(.)) :\leftrightarrow \mathsf{val}(x,y) = z \qquad \text{Cfr. §2.3.2;}$$
$$F(D) :\leftrightarrow \mathsf{val}(x,y) = z$$
$$F(\mathsf{Sat}) :\leftrightarrow \mathsf{Tr}_{\mathsf{U}}(x[y]^{\bullet})$$
τ_2 commutes with propositional connectives;
$$(\forall v_i^s \varphi)^{\tau_2} :\leftrightarrow \forall x((\varphi)^{\tau_2})$$
$$(\forall v_i^o \varphi)^{\tau_2} :\leftrightarrow \forall x(\delta^o(x) \rightarrow (\varphi)^{\tau_2})$$
$$(\forall v_i^{sq} \varphi)^{\tau_2} :\leftrightarrow \forall x(\delta^{sq}(x) \rightarrow (\varphi)^{\tau_2})$$

Table 4.3: The relevant details of the translation τ_2.

be used to relativise quantifiers of $\mathcal{L}_{\mathsf{U}}$.[23] Thus we map objects belonging to the 'mathematical' component of $\mathsf{CTD}^*[\mathsf{U}]$ to new constants in C. The domain of quantifiers ranging over 'sequences' will consist in sequences of these constants. Let $\mathsf{Seq}(x)$ be the Δ_1—provably in $I\Sigma_1$—predicate of the language of arithmetic expressing that x codes a sequence , and $\mathsf{lh}(x)$ the Δ_1—provably in $I\Sigma_1$—function in $\mathcal{L}_{\mathsf{PA}}$ that maps the sequence x to its length, and $(x)_i$ the functional expression of $\mathcal{L}_{\mathsf{PA}}$ indicating the i^{th} element of the sequence x,[24] we let

$$\delta^{sq}(x) :\leftrightarrow \mathsf{Seq}(x) \wedge \forall i < \mathsf{lh}(x)(\delta^o((x)_i)).$$

Since we translate the 'syntactic' part of the language of $\mathcal{L}_{\mathsf{TD}}$ by means of the Ackermann interpretation (see §2.3.1), there is no need to specify a domain for the quantifiers of $\mathcal{L}_{\mathsf{TD}}$ ranging over hereditarily finite sets.

[23] For the sake of readability, we leave aside the complication given by the presence of the equality in the language, but it should be obvious how to proceed.

[24] We are very rough here, $\mathsf{Seq}(x)$ and $\mathsf{lh}(x)$ are in fact $\Sigma_0^{exp}(exp)$ in $I\Delta_0(exp)$.

Remark 1. Before completing the proof, we highlight a crucial point: in the proof of Lemma 4.4, we briefly hinted at the fact that the finite subsystem of HF formalising the syntax for the finite subsystem U of ZF was interpreted in the language of set theory in such a way that if a $\mathcal{L}_{\mathsf{HF}}$-formula $\varphi(k_1, ..., k_n)$ defined in HF a syntactic relation, then the formula $(\varphi)^{\tau_1}(k_1, ..., k_n)$ of $\mathcal{L}_{\in}$ would represent the same syntactic relation in ZF. This can be easily obtained in a set theoretic context, given the translation defined in Table 4.2.

This assumption, however, when $\mathcal{L}_{\mathsf{HF}}$ is translated into the language of arithmetic, needs at least some additional remark. Intuitively, it seems that we would need to code the syntax of the system U in PA in a suitable way.

Given the synonymity of HF and PA, however, we can ensure this point by reasoning within HF already. We first recall that the Inverse Ackermann interpretation $\mathfrak{e}(.)$ of finite sets into natural numbers can be defined within HF by recursion on rank, so that

$$\mathfrak{e}(k) = \begin{cases} 0, & \text{if } k = 0; \\ 2^{\mathfrak{e}(k_1)} + ... + 2^{\mathfrak{e}(k_n)}, & \text{if } k = \{k_1, ..., k_n\}. \end{cases} \tag{4.9}$$

From §2.3.1, as $I\Delta_0(\mathsf{exp})$ is a subtheory of HF, the Ackermann interpretation can be realised in HF. A fortiori, for every formula $\varphi(k_1, ..., k_n)$ of $\mathcal{L}_{\mathsf{HF}}$ we can prove that there is a formula $(\varphi)^*$ of $\mathcal{L}_{\mathsf{HF}}$—where $(.)^*$ is the translation induced by $\mathfrak{e}(.)$—such that

$$\mathsf{HF} \vdash \forall k_1, ..., k_n(\varphi(k_1, ..., k_n) \leftrightarrow (\varphi)^*(\mathfrak{e}(k_1), ..., \mathfrak{e}(k_n)).$$

Thus for every sentence of $\mathcal{L}_{\mathsf{HF}}$ there is a sentence σ^* of $\mathcal{L}_O$ such that

$$\mathsf{HF} \vdash \sigma \leftrightarrow \sigma^*.$$

Reasoning within HF, we thus have that if $\varphi(k_1, ..., k_n)$ expresses a syntactic relation about O between the sets $k_1, ..., k_n$, then the formula $(\varphi)^*(\mathfrak{e}(k_1), ..., \mathfrak{e}(k_n))$ can be taken to represent the same syntactic relation between the finite ordinals $\mathfrak{e}(k_1), \ldots, \mathfrak{e}(k_n)$. The formula $\mathsf{Cterm}^s_{\mathcal{L}_\mathsf{U}}(k)$, for instance, expressing in HF that k is a closed term of the language of arithmetic, would be translated as $(\mathsf{Cterm}^s_{\mathcal{L}_\mathsf{U}})^*(\mathfrak{e}(k))$. Our assumption will be that $(\mathsf{Cterm}^s_{\mathcal{L}_\mathsf{U}})^*(\mathfrak{e}(k))$ expresses in HF already, and thus in PA, that $\mathfrak{e}(k)$ represents a closed term of $\mathcal{L}_\mathsf{U}$. We will thus write $(\mathsf{Cterm}^s_{\mathcal{L}_\mathsf{U}})^*(\mathfrak{e}(k))$ as $\mathsf{Cterm}^o_{\mathcal{L}_\mathsf{O}}(x)$. ▪

Proof of Lemma 4.5 continued. τ_2 also associates to the function symbol $.(.)$ of type $(sq, s) \to o$ and to its generalisation D the Δ_1 arithmetical evaluation function $\mathsf{val}(x, y) = z$, already introduced in Chapter 2,[25] which will then apply to a (code of) a $\mathcal{L}_\mathsf{U}$-term with free variables and to a sequence in the sense of $\delta^{sq}(.)$. Moreover, to Sat of sort (sq, s) τ_2 assigns the formula $\mathsf{Tr}_\mathsf{U}(z)$ introduced above, which applies, given the code of a formula φ of $\mathcal{L}_\mathsf{U}$ whose only free variables are coded by the sequence $\vec{v}$, to its counterpart $\overline{\varphi}[\vec{v}]^\bullet$ such that $\mathsf{cFml}_{\mathcal{L}_\mathsf{U}}(\overline{\varphi}, \vec{v}, \overline{\varphi}[\vec{v}]^\bullet)$.

We now show that τ_2 is indeed a relative interpretation of $\mathsf{CTD}^*[\mathsf{U}]$ into PA. Axioms of U are straightforwardly satisfied, and so are the basic axioms of HF before induction. The axioms for denotation are captured by the properties of the evaluation function, modulo *Remark 1*. The translations of the axioms for satisfaction are also satisfied by PA. For instance, to see that PA proves the translation of

$$\forall a \forall k, l(\mathsf{Cterm}^s_{\mathcal{L}_\mathsf{O}}(k) \wedge \mathsf{Cterm}^s_{\mathcal{L}_\mathsf{O}}(l) \to (\mathsf{Sat}(a, k \underset{\cdot}{=} l) \leftrightarrow D(a, k) = D(a, l))),$$

we notice that if $\mathsf{Tr}_\mathsf{U}(x)$, where x is equal to a term $s \underset{\cdot}{=} t[z]^\bullet$ corresponding to an equality between closed terms of the language of arithmetic coded by $s[z]^\circ$ and $t[z]^\circ$ respectively, then

$$\mathsf{Tr}_\mathsf{U}(s \underset{\cdot}{=} t[z]^\bullet) \leftrightarrow \mathsf{val}(s, z) = \mathsf{val}(t, z).$$

Therefore, modulo a suitable formalisation of the syntax of U in PA, the translations of all

[25] See also [Hájek&Pudlák 1993, p. 53].

the other axioms for satisfaction are provable in $\mathsf{PA} + \mathsf{Mod}_{\mathsf{U}}$.

For the extended induction $\mathsf{HF3}^*$ we notice that the induction of PA is more than what is needed to perform induction on formulas such as $\mathsf{Tr}_{\mathsf{U}}(x)$.[26] *qed.*

By applying Fact 4.1 and Fact 4.3, we then obtain

Corollary 4.3. $\mathsf{CTD}[\mathsf{PA}] \leqslant \mathsf{PA}$.

We now show that the interpretation just defined in Lemma 4.5 allows us to highlight an important difference between CTD[PA] and CT—specular to what we detected already in the case of CTD[ZF] and CT[ZF]. Without interactions between the syntactic schema proper of the theory of truth and the mathematical schemata of the object theory, the inductive reasoning leading to the establishment of the soundness of PA, available in CT, is not at our disposal.

Theorem 4.2. $\mathsf{CTD}[\mathsf{PA}] \nvdash \mathsf{Con}^s_{\mathsf{PA}}$.

Proof. Assume $\mathsf{CTD}[\mathsf{PA}] \vdash \mathsf{Con}^s_{\mathsf{PA}}$. Then $\mathsf{CTD}^*[\mathsf{U}] \vdash \mathsf{Con}^s_{\mathsf{PA}}$ with $\mathsf{CTD}^*[\mathsf{U}]$ as in Lemma 4.5. Thus $\mathsf{PA} + \mathsf{Mod}_{\mathsf{U}} \vdash (\mathsf{Con}^s_{\mathsf{PA}})^{\tau_2}$. But for any finite U, PA proves $\mathsf{Con}_{\mathsf{U}}$, and thus by Fact 4.3, $\mathsf{Mod}_{\mathsf{U}}$. Thus $\mathsf{PA} \vdash (\mathsf{Con}^s_{\mathsf{PA}})^{\tau_2}$. What remains to be done is to make sure that $(\mathsf{Con}^s_{\mathsf{PA}})^{\tau_2}$ is not provable in PA, so that it can be considered a 'canonical' consistency statement for PA. Following what has been already said in *Remark 1* and done in the proof of Theorem 4.1, we argue that it is possible to formalise the syntax of U in PA in such a way that $\mathsf{Con}^s_{\mathsf{PA}}$, modulo Ackermann interpretation and its inverse, is indeed a canonical consistency statement for PA. In other words, by employing the relationships between HF and PA explored in §2.3.1, we can argue as follows in a more direct way: $\mathsf{PA} \trianglelefteq \mathsf{HF}$ (see. §2.3.1.), so that we have

[26] Presumably Δ_2-induction suffices.

$$
\begin{aligned}
&\mathsf{HF} \vdash (\mathsf{Con}^s_{\mathsf{PA}})^{\tau_2} && \text{since } \mathsf{PA} \subseteq \mathsf{HF}; \\
&\mathsf{HF} \vdash (\mathsf{Con}^s_{\mathsf{PA}})^{\tau_2} \leftrightarrow \mathsf{Con}^s_{\mathsf{PA}} && \text{by } \textit{Remark } 1 \\
&\mathsf{HF} \vdash \mathsf{Con}^s_{\mathsf{PA}}.
\end{aligned}
$$

Since HF is synonymous with PA, the last line contradicts Gödel's second Incompleteness theorem. *qed.*

By Modus Tollens, we have

Corollary 4.4. $\mathsf{CTD}[\mathsf{PA}] \nvdash \mathsf{GRP}^s_{\mathsf{PA}}$.

As one might expect, the possibility of performing syntactic induction on the length of the proofs in O, when O is schematically axiomatised, to obtain $\mathsf{GRP}^s_{\mathsf{O}}$ is restored if we get around the limitations induced by the separation between syntax and object theory. In particular, by letting

$$\mathsf{AxT}^s_{\mathsf{O}} :\leftrightarrow \forall a \forall k(\mathsf{Ax}^s_{\mathsf{O}}(k) \to \mathsf{Sat}(a,k)),$$

we have

Corollary 4.5. For any choice of O, $\mathsf{CTD}[\mathsf{O}] + \mathsf{AxT}^s_{\mathsf{O}} \vdash \mathsf{GRP}^s_{\mathsf{O}}$.

$\mathsf{AxT}^s_{\mathsf{O}}$ is exactly what is required to overcome the limitations given by Corollary 4.4 and Corollary 4.2. In CTD[O] then, when O is a schematically axiomatised theory,[27] $\mathsf{AxT}^s_{\mathsf{O}}$ is a quite powerful principle. In particular, if one considers CTD[PA] an accurate axiomatisation of the Taskian metatheory at least as CT is, it is apparent that CT proves the Global Reflection Principle for PA only for a quite fortunate coincidence.

[27] We considered the case of PA and ZF, but we can apply these considerations to the other schematically axiomatised theories.

4.5 Conservativity

We will later characterise the deflationary standpoint about truth by means of three general tenets (see Chapter 6): the truth predicate expresses no property; the truth predicate is exclusively a logico-linguistic device for disquotation and a non dispensable tool for expressing generalisations; the notion of truth does not play a substantial role in scientific explanations.

Horsten, Shapiro and Ketland in [Horsten 1995, Shapiro 1998, Ketland 1999] have translated this alleged insubstantiality of the truth predicate into the claim that a formal theory of truth constructed over an object theory O, in order to be faithful to the deflationary spirit, has to be conservative over the object theory T.[28] Halbach in [Halbach 2001a] has shown that if O is pure predicate logic (FOL), even a simple axiomatisation of the truth predicate based on the T-schema, call it $\mathsf{TB}[\mathsf{FOL}]$, is nonconservative over FOL, as it proves the existence of at least two objects, whereas pure predicate logic can only prove the existence of a single object.[29]

To size the impact that the so-called ‘conservativeness argument’ has on deflationism, it seems thus appropriate to focus on the conservativity of the theory of truth over the base theory O. We will later see that the very notion of *conservativity of the theory of truth over the base theory*, when theories of truth with ‘disentangled syntax’ are concerned, needs further clarification. We now introduce some useful definitions.

Definition 4.4 (Expansion of a model)**.** Let $\mathcal{M} = (|\mathcal{M}|, \mathcal{I})$ be a model of the first-order language $\mathcal{L}$. Moreover, let $\mathcal{L}' = \mathcal{L} \cup X$, where X is a set of new nonlogical symbols not in $\mathcal{L}$ already. We call the resulting $\mathcal{L}'$-structure $\mathcal{M}' = (|\mathcal{M}|, \mathcal{I} \cup \mathcal{I}')$, where $\mathcal{I}'$ is an interpretation of the symbols in X, an *expansion* of $\mathcal{M}$. Symmetrically, $\mathcal{M}$ is called the *reduct* of $\mathcal{M}'$.

Let $\mathsf{CT}{\restriction}$ be the theory, introduced in §2.3.2 already, obtained from PA in $\mathcal{L}_{\mathsf{PA}} \cup \{T\}$ by

[28] For the notion of conservative extension, see Definition 4.6.

[29] Halbach considers this as a compelling evidence supporting the thesis that the truth predicate is not a *purely logical* device, as theorizing with the help of a truth predicate is always relative to a given ontology.

adding to it the Tarskian compositional clauses turned into axioms and by restricting the induction of PA to arithmetical formulas.

Definition 4.5 (Satisfaction Class for a model of PA)**.** Let $\mathcal{M}$ be a model of PA. A set $S \subseteq |\mathcal{M}|$ is a (full) satisfaction class if and only if $(\mathcal{M}, S)$ models CT$\restriction$.

Definition 4.6 (Conservativity)**.**

(i) Let $\mathcal{L}_{\mathsf{S}}$ be a language and Γ a class of formulas of $\mathcal{L}_{\mathsf{S}}$. A system T in $\mathcal{L}_{\mathsf{T}} \supseteq \mathcal{L}_{\mathsf{S}}$ is *proof-theoretically* Γ*-conservative* over the system S if and only if each Γ-formula φ provable in T is also provable in S. We say that T is a *conservative extension* of S if Γ coincides with the class of all $\mathcal{L}_{\mathsf{S}}$-formulas.

(ii) (Craig and Vaught). A theory T in $\mathcal{L}_{\mathsf{T}}$ is *model-theoretically conservative* over a system S in $\mathcal{L}_{\mathsf{S}}$ with $\mathcal{L}_{\mathsf{S}} \subseteq \mathcal{L}_{\mathsf{T}}$ if and only if any model of S can be expanded to a model of T.

Model-theoretic conservativity implies proof-theoretic conservativity, but the converse does not always hold.[30] A well-known example of a situation in which proof-theoretic conservativity holds whereas model-theoretic conservativity fails concerns the theory CT$\restriction$. By Lachlan's Theorem,[31] not every model of PA can be expanded to a model of CT$\restriction$, as only recursively saturated models of PA have a full satisfaction class. But every consistent extension of PA has a model which is not recursively saturated. Thus CT$\restriction$ is not model-theoretically conservativite over PA. However, CT$\restriction$ is known to be proof-theoretically conservative over PA.[32]

An important feature of the theory CTD[O] is expressed by the following result:

Theorem 4.3. CTD[O] is model-theoretically conservative over O.

[30] We also notice that a theory T is conservative over S if and only if there is an extension $\mathcal{M}' \models \mathsf{T}$ of a model $\mathcal{M} \models \mathsf{S}$ such that the reduct of $\mathcal{M}'$ to $\mathcal{L}_{\mathsf{S}}$ is an elementary extension of $\mathcal{M}$.

[31] Cfr. [Halbach 2011, pp. 89ff] and [Lachlan 1981, Kaye 1991].

[32] For the standard result, see [Halbach 2011, Kotlarsky et alii 1981, Kaye 1991]. For new proofs, cfr. [Enayat&Visser 2013] and [Leigh 2013]. Another example is represented of a theory of truth which is proof-theoretically but not model-theoretically conservative over PA is the theory UTB. In [Leigh&Nicolai 2013, p. 14, footnote 21] a simpler example, not involving axioms for truth, is considered.

The key detail of the proof, which also makes the disentangled setting different from the usual one, is represented by the possibility of employing a standard interpretation of the syntax and of the length of the sequences of variable assignments—governed by SQ—*by stipulation*, even in the presence of a non-standard model of O.

Remark 2. We consider the case of a nonstandard $\mathcal{M} \vDash \mathsf{PA}$.[33] The following considerations, however, can easily be applied to an ω-nonstandard model of ZF.

It is well-known—the so-called Overspill Principle—that for any formula $\varphi(x, y)$ of the language of arithmetic, if

$$\mathcal{M} \vDash \varphi(\overline{n}, \overline{m})$$

for any $n \in \omega$ and $m \in |\mathcal{M}|$, then there is a nonstandard $c \in |\mathcal{M}|$ such that

$$\mathcal{M} \vDash \forall x \leq \overline{c}\ \varphi(x, \overline{m}).$$

Therefore if $\mathcal{M} \vDash \varphi(\overline{n}, \overline{m})$ for infinitely many $n \in \omega$, then there is also a nonstandard $c \in |\mathcal{M}|$ such that $\mathcal{M} \vDash \varphi(\overline{c}, \overline{m})$. As a result, any nonstandard model of PA will contain nontandard (codes of) variables, terms, formulas, sentences and so on for the representations of other relevant syntactic and logical notions.

Furthermore, if $\mathcal{M}$ is nonstandard, and $(\mathcal{M}, S)$ is like in Definition 4.5, by the compositional axiom for negation and by the previous considerations we conclude that there will always be a nonstandard sentence in our satisfaction class. This is why the process of defining a satisfaction class for a nonstandard model of PA, in the absence of the extended induction, becomes quite convoluted.[34]

What makes theories of truth with 'disentangled syntax' such as CTD[PA] special in this respect—although somewhat uninteresting from a strictly mathematical point of view—is

[33] We always identify, as in [Halbach 2011], the nonstandard $\mathcal{M}$ with its expansion including the interpretation of constants $\overline{a}$ for every $a \in |\mathcal{M}|$ added to the initial language.

[34] See [Kotlarsky et alii 1981] and [Enayat&Visser 2013].

that despite the fact that we still don't have the induction schema of PA extended to the full language $\mathcal{L}_{\mathsf{TD}}$, for the definition of a satisfaction class for any model $\mathcal{M}$ of PA we can safely assume that there are no nonstandard syntactic entities.

Proof of Theorem 4.3. Given an arbitrary model $\mathcal{M}$ of O, we construct its expansion $\mathcal{D} = (|\mathcal{M}|, |\mathcal{M}|^{\omega}, \mathbb{H}, \mathcal{I}^{o} \cup \mathcal{I}^{+})$, where $\mathbb{H}$ is the domain of the standard model of HF defined in §2.3, $|\mathcal{M}|^{\omega}$ is the domain of variables of sort sq and $\mathcal{I}^{+}$ specifies the interpretation of the new vocabulary of $\mathcal{L}_{\mathsf{TD}}$. In particular, the function $.(.)$—and consequently D, if function symbols are present in $\mathcal{L}_{O}$—is naturally interpreted as a function

$$f^{\mathcal{D}}\colon |\mathcal{M}|^{<\omega} \times \mathbb{H} \to |\mathcal{M}|$$

from the cartesian product of the set of finite sequences of $\mathcal{M}$-objects and of the standard model of arithmetic into the domain of $\mathcal{M}$. Finally, the extension S of the satisfaction predicate is defined standardly as:

$$S = \{(x, y) : y = \sharp\,\varphi \text{ and } \sharp\,\varphi \in \mathbb{H} \text{ for some } \varphi \text{ of } \mathcal{L}_{O}, x \in |\mathcal{M}|^{<\omega} \text{ and } \mathcal{M} \vDash \varphi[x]\} \tag{4.10}$$

where $\sharp\,\varphi$ denotes the element of $\mathbb{G}$ coding the formula φ of $\mathcal{L}_{O}$. Therefore S is taken to be the set of pairs (x, y) where x is a sequence of objects—of standard length—in the sense of the domain of an arbitrary model $\mathcal{M}$ of the object theory O, y is a finite set in $\mathbb{G}$ coding a $\mathcal{L}_{\mathsf{O}}$-formula φ, and φ is satisfied by $\mathcal{M}$ according to the variable assignment x. A complete definition of S would have to be carried out on the complexity of φ, but this creates no problems as syntactic objects are well-founded. We have thus constructed the expansion $\mathcal{D}$ of $\mathcal{M}$.

We then show that $\mathcal{D}$ is indeed a model of CTD[O]. Axioms of O and of HF are clearly satisfied by $\mathcal{D}$. SQ and HF3* satisfied by $\mathcal{D}$ by a straightforward inductive argument in the metatheory on the (standard) length of sequences and syntactic objects respectively. Axioms for satisfaction are satisfied by $\mathcal{D}$ by definition of S. *qed.*

Corollary 4.6. $\mathsf{CTD}[\mathsf{O}]$ is proof-theoretically conservative over O.

Corollary 4.7. $\mathsf{TB}[\mathsf{O}]$ and $\mathsf{UTB}[\mathsf{O}]$ are model-theoretically and thus proof-theoretically conservative over O.

By the same argument employed in Theorem 4.3, we have also:

Corollary 4.8. $\mathsf{CTD}[\mathsf{O}] + \mathsf{AxT}^{s}_{\mathsf{O}}$ is model-theoretically, and thus proof-theoretically conservative over the schematic theory O.

4.5.1 Extending Schemata.

In the present subsection we try to answer to the following question: what happens, in the case in which O is schematically axiomatised, if we allow the arbitrary formulas of $\mathcal{L}_{\mathsf{TD}}$ to appear into the schemata of O? Some philosopher might in fact judge the restriction imposed on schemata of O, as well as the negative emphasis put on the interactions between syntactic and mathematical schemata, somewhat arbitrary and yet unjustified, given the benefits that a schematic formulation of mathematical theories displays in terms of proof theoretic strength, ontological parsimony and expressive power.[35]

We notice that the definition of schematically axiomatised theories given in Chapter 2 already rules out arbitrary extensions of schemata of the object theory, at least when the Tarskian picture is at issue. Moreover, we expect that the motivational remarks contained in Chapter 2 are also sufficient to defend the thesis that in this particular construction the interaction between the two roles of schemata can be judged at least confusing. At any rate, the results that will be shortly presented show how important is the restriction of the schemata of O to obtain a conservative truth theory.

[35] See for instance [Feferman 1991] and [McGee 1997].

It is important to recall, at this stage, our indexing conventions of terms and formulas of $\mathcal{L}_{\mathsf{TD}}$. Since we are now formalising syntactic operations and notions on and about $\mathcal{L}_{\mathsf{PA}}$ within PA itself, we index them with the superscript o and we place a line over codes of $\mathcal{L}$-terms instead of putting them into Gödel corners. Let us consider the following schemata:

(Ind$^+$) $\quad F(0^o) \wedge \forall x(F(x) \to F(S^o x)) \to \forall x(F(x))$
for all formulas F of $\mathcal{L}_{\mathsf{TD}}$ with x in V_o;

(Sep$^+$) $\quad \forall x \exists y \forall z(z \in y \leftrightarrow z \in x \wedge F(z))$
for any formula F of $\mathcal{L}_{\mathsf{TD}}$ and x, y, z in V_o;

(Rep$^+$) $\quad \forall x\big[\forall u \in x \exists! v F(u,v) \to \exists y \forall v\big(v \in y \leftrightarrow (\exists u \in x)F(u,v)\big)\big]$
for any formula F of $\mathcal{L}_{\mathsf{TD}}$ and u, v, x, y in V_o.

(HF3$^+$) $\quad F(0^o) \wedge \forall x, y(F(x) \wedge F(y) \to F(x \triangleleft^o y)) \to \forall x\, F(x)$
for any formula F of $\mathcal{L}_{\mathsf{TD}}$ and x, y in V_o.

And the resulting theories:

$$\begin{aligned}
\mathsf{CTD[PA]}^+ &:= \mathsf{CTD[PA]} + \mathsf{Ind}^+\\
\mathsf{CTD[ZF]}^+ &:= \mathsf{CTD[ZF]} + \mathsf{Sep}^+ + \mathsf{Rep}^+\\
\mathsf{CTD[HF]}^+ &:= \mathsf{CTD[HF]} + \mathsf{HF3}^+.
\end{aligned}$$

We introduce some notions that will be employed in the proof of our next result. Working within PA, we start with a suitable partial truth predicate $\mathsf{Tr}^o_{\Sigma,n}(x)$ which applies to codes of sentences of the language $\mathcal{L}_{\mathsf{PA}}$ represented in $\mathcal{L}_{\mathsf{PA}}$ itself. Following [Hájek&Pudlák 1993, § I.1.(d)], we define it, for every Σ_n-formula $\varphi(v_1, \ldots, v_n)$ of $\mathcal{L}_{\mathsf{PA}}$ and for each $n \geq 0$, in terms of a suitable partial satisfaction predicate as

$$\mathsf{Tr}^o_{\Sigma,n}(\ulcorner\varphi\urcorner[\dot{v}_1/\ulcorner v_1\urcorner,\dots,\dot{v}_n/\ulcorner v_n\urcorner]) :\leftrightarrow \mathsf{Sat}^o_{\Sigma,n}(\langle\rangle,\ulcorner\varphi\urcorner[\dot{v}_1/\ulcorner v_1\urcorner,\dots,\dot{v}_n/\ulcorner v_n\urcorner])$$

where $\langle\rangle$ denotes the empty sequence,[36] and $\dot{x}$ the usual function that maps each number to its numeral, so that still for every Σ_n-formula $\varphi(v_1,\dots,v_n)$ and for each $n \geq 0$

$$\mathsf{I}\Sigma_1 \vdash \mathsf{Tr}^o_{\Sigma,n}(\overline{\varphi}[\dot{v}_1/\ulcorner v_1\urcorner,\dots,\dot{v}_n/\ulcorner v_n\urcorner]) \leftrightarrow \varphi(v_1,\dots,v_n).$$

Theorem 4.4. Let O be either PA, ZF of HF. Then $\mathsf{CTD}[\mathsf{O}]^+$ is a nonconservative extension of O.

Proof. We present the argument for $\mathsf{CTD}[\mathsf{PA}]^+$. The strategies for $\mathsf{CTD}[\mathsf{ZF}]^+$ and $\mathsf{CTD}[\mathsf{HF}]^+$ are similar.

We claim that $\mathsf{CTD}[\mathsf{PA}]^+$ proves

$$\forall x\big(\mathsf{Sent}^o_{\mathcal{L}_{\mathsf{PA}}}(x) \land \mathsf{Bew}^o_{\mathsf{PA}}(x) \rightarrow \forall a\forall k\forall n(a(k) = x \rightarrow \exists y(a(n) = y \land \mathsf{Sat}(a,\ulcorner\mathsf{Tr}^o_{\Sigma,v_n}(v_k)\urcorner)))\big). \tag{4.11}$$

(4.11) is provable by induction on the length of the proof in PA, with a subsidiary induction to establish the base case, both of which are available in $\mathsf{CTD}[\mathsf{PA}]^+$. In the proof of (4.11), there are two non trivial cases in which Lemma 4.1—the Regularity Lemma—plays an important role. The first of these concerns the logical axioms of PA. The second concerns the induction axioms of PA. Given an arbitrary instance of induction, x, we require to show that

$$\forall a\forall k\forall n(a(k) = x \rightarrow \exists y(a(n) = y \land \mathsf{Sat}(a,\ulcorner\mathsf{Tr}^o_{\Sigma,v_n}(v_k)\urcorner))). \tag{4.12}$$

[36] Definable already in $I\Delta_0(\exp)$.

This follows, however, from the application of Ind^{+} to the formula:

$$\psi(z) := \forall a \forall i, j, n \exists y (a(n) = y \wedge \mathsf{Fml}^{o}_{\mathcal{L}_{\mathrm{PA}}}(a(i)) \wedge a(j) = z \wedge \mathsf{Sat}(a, \ulcorner \mathsf{Tr}^{o}_{\Sigma, v_n}(\mathsf{sub}^{o}(v_i, \overline{v}, v_j)) \urcorner)). \tag{4.13}$$

By instantiating a refutable sentence ξ in O within (4.11), we obtain $\neg\mathsf{Bew}^{o}_{\mathsf{PA}}(\overline{\xi})$, i.e. $\mathsf{Con}^{o}_{\mathsf{PA}}$.

qed.

To the challenge mounted by Horsten, Shapiro and Ketland to the deflationist's tenet of the insubstantiality of truth, Field in [Field 1999] has answered by highlighting how the nonconservativity of theories in the style of CT is deeply tied with an increase in the mathematical resources of the theory. In particular, in the case of CT, the addition of mathematical power is bound to the extension of the induction schema of PA to the entire vocabulary of the theory of truth. At any rate, the characterisation of what 'mathematical power' could mean has not been made entirely clear by Field. In fact, as we have seen following Heck's analysis of the two roles of the induction schema in the proof of $\mathsf{GRP}_{\mathsf{PA}}$, the extended induction is more than a mere mathematical principle.

We will discuss these issues more extensively in Chapter 6. However, there is a sense in which we can already grant Field's analysis of having hit a crucial point. The theory $\mathsf{CTD}[\mathsf{O}]$ might appear to be somewhat unnatural, given the strict separation between syntax an mathematics. However, it also provides us with a clear picture of which roles are assigned to the schemata needed for the establishment of metamathematical arguments *on* the system O involving the notion of truth on the one hand, and schemata needed for the development of the 'subject matter' of O on the other. What Theorem 4.4 tells us then is that in $\mathsf{CTD}[\mathsf{O}]$, with O schematically axiomatised, we also have a clear picture of what 'increasing the mathematical power' of Peano Arithmetic, in Field's sense, might mean; that is, to allow schemata of the theory of truth to conclude the truth of formulas at each substitutional instance, as it happens in (4.13), into metamathematical arguments *on* the system O. And of course, if we 'increase the mathematical power' of O in the sense just described, we *do* get a nonconservative extension of O assuming a compositional theory of truth over it.

5 A Portrait of Metamathematical Practice

If approached from a different perspective, the project of the 'disentanglement' of the syntax from the object theory in the construction of axiomatic theories of truth may be seen as a first step towards a complete formalisation of our informal metamathematical practice. In the present chapter we try to fully accomplish this task and offer, by means of the framework already introduced in the previous chapters, a faithful picture of the all-present interplay between formal and informal metatheoretic claims. In particular, we will first decompose our metamathematical practice in different categories; we will then introduce a two-sorted theory obtained by adding suitable axioms relating entities belonging to the domain of the theory formalising the syntax for the object theory and elements of the domain of the object theory itself; we will finally extend this theory with axioms for denotation and truth and argue that the resulting theory models the interactions between the different components of our metamathematical reasoning.

5.1 Aspects of Metamathematical Reflection: Informal Morphology$^{\mathsf{O}}$

Metamathematical reflection is essentially stratified. A clear distinction between informal and formal metatheoretic reasoning can easily be detected. Let us consider a unary function symbol f belonging to the language of our object theory O and let $\mathfrak{t}$ be a term of $\mathcal{L}_{\mathsf{O}}$. According to the rules of formation for terms proper of the grammar of $\mathcal{L}_{\mathsf{O}}$, there is an operation $\check{f}$ defined on terms of $\mathcal{L}_{\mathsf{O}}$ which unambiguously yields a well-formed functional expression of $\mathcal{L}_{\mathsf{O}}$ denoting the application of f to $\mathfrak{t}$. $\check{f}$ belongs to what we will call, borrowing a little from the Tarskian jargon, Informal Morphology$^{\mathsf{O}}$.[1] This is the area of our metatheoretic reflection, living in informal mathematics, to which also the characterisation of the informal development of the syntax of O, carried out in Section 2.2, belongs: there we state and prove the unique readability lemma, there we define classes of structured strings of symbols which correspond to the syntactic categories of the language under consideration; this 'theory' is usually also taken to be strong enough to establish the informal statement expressing our belief in the consistency of O.

The operation $\check{f}$ is usually defined in terms of a more general (binary) total operation, call it $\mathsf{app}(x,y)$, which enables one to take two expressions of $\mathcal{L}_{\mathsf{O}}$ and construct another expression of $\mathcal{L}_{\mathsf{O}}$. By repeated application of $\mathsf{app}(x,y)$, we can construct arbitrary strings of symbols from the alphabet of $\mathcal{L}_{\mathsf{O}}$. It does not matter whether $\mathsf{app}(x,y)$ is understood as a concatenation operation or as a pairing function: its characterisation still belongs to what we called Informal Morphology$^{\mathsf{O}}$. In a word, the system O *qua* formal system, that is a set of syntactic entities and relations among them, represents the scope of Informal Morphology$^{\mathsf{O}}$. One remarkable aspect of the informal theory just depicted—which is missing in all the usual attempts to formalise Informal Morphology$^{\mathsf{O}}$—is that its objects seem to possess

[1] To give a well-known analogy, our Informal Morphology$^{\mathsf{O}}$ is what is called 'Syntax' for the case of Peano Arithmetic in [Boolos 1993, pp. 33ff].

a genuine syntactic character. An attempted formalisation, therefore, should avoid—if possible—any reduction of these objects to other entities such as numbers or set: *no coding ought to be involved*. In general in fact,

> Identifying numbers and expressions is a notational simplification at best, but in informal metatheoretic discussion the theory of syntax and the theory of natural numbers should be kept separate: expressions are not numbers. ([Halbach 2011, p. 316])

Remark 4.1. We just touched on a point which, for reasons of time and space, would deserve more attention than what it receives in the present work: a faithful formalisation of the informal morphology for O would be better carried out in a theory which takes syntactic objects as primitives and does not reduce them to other sorts of entities, such as finite sets. There are various degrees in which the elimination of the coding can be carried out.

A first attempt consists in developing theories which take primitive expressions of $\mathcal{L}_{\mathsf{O}}$ as constants in their language: for instance, to stick with an example that should be familiar to the reader, we might consider HF plus atoms for $c, v, f, R, \neg, \rightarrow, \forall$: a theory which we can call HFA. Atoms of course will not be sets, and we may stipulate that 0—that is the constant for the empty set in HF—is not an atom as well, so that ordinals cannot contain any urelements. Extensionality has to be provably restricted to finite sets. It is relatively straightforward to show that the properties of HF described in Chapter 3 are enjoyed by HFA as well.[2] Moreover, HF and HFA are clearly mutually interpretable and it may be conjectured that the two theories are synonymous, that is they have a common definitional extension. When applied to the formalisation of Informal Morphology$^{\mathsf{O}}$, we notice that HFA eliminates the arbitrariness related to the assignment of closed terms of $\mathcal{L}_{\mathsf{HF}}$ to primitive symbols of $\mathcal{L}_{\mathsf{O}}$ carried out in §3.2. However, it can still be argued, and rightly so, that coding has not been eliminated yet, as syntactic entities, already at the level of complex expressions, are understood as finite sets.

[2] We have partially verified this in detail in an unpublished note.

A second option may consists in considering direct axiomatisations of concatenation. Besides the approach presented in [Grzegorczyck 2005], [Grzegorczyck&Zdanowski 2007] and [Visser 2008a], which pointed mainly at a proof of the undecidability of the resulting theory, [Halbach&Leigh 2013] consider a weak theory of concatenation A with the empty string and primitive quotation and substitution functions, essentially motivated by an elimination of the usual reduction of expressions to other entities. The quotation function, which associates to any expression e a quotation constant $\overline{e}$, together with the concatenation function gives the means to talk about strings of the language $\mathcal{L}_{\mathsf{A}}$ of A within A itself; the primitive substitution function, on the other hand, makes the route to diagonalisation and the derivation of logical antinomies surprisingly uncomplicated. If the target is a formalisation of the morphology of a first-order theory, however, and A itself among them, A would have to be strengthened with some induction on expressions[3] to extract a good notion of finite sequence and define the relevant syntactic and logical categories. Although somewhat preferable, however, none of these two options will be considered below; we will stick in fact with the theory HF. ▪

For a sufficiently strong object theory, the proofs of the diagonal lemma, of the incompleteness theorems and related paradoxes call for an 'interpretation' of what we called $\mathsf{Informal\ Morphology}^{\mathsf{O}}$ into the object theory O itself. Since we are dealing with a portion of informal mathematics, the notion of relative interpretation introduced in the previous chapter cannot immediately help; at any rate, the aim is to set a correspondence between the domain of discourse of our informal morphology, that is a universe of syntactic objects, and objects in the sense of the intended domain of our object theory.

The usual procedure involves the following steps:[4] (i) we fix a portion of informal mathematical language which acts as language of $\mathsf{Informal\ Morphology}^{\mathsf{O}}$; (ii) we define a translation of this language into the language of the object theory; (iii) we verify that the translation is

[3] Presumably bounded induction on strings is sufficient.
[4] Cfr. [Feferman 1960, Feferman 1991], [Smorynski 1977] and [Boolos 1993].

well-behaved, namely we ensure that a formula of the language of the object theory which is the translation of a predicate of Informal Morphology$^{\mathsf{O}}$ is true only of those objects that are the counterparts of the syntactic objects of which the syntactic predicate holds. As a consequence, (iv) a sentence of the language of the object theory expressing a statement about its syntax will be provable in the object theory *only if* its counterpart is provable in the syntax.

[Boolos 1993] notices that it is not possible, for a suitable O, to render step (iv) an *if and only if* clause, so that the provability in the object theory of a sentence p of the language of the object theory expressing a syntactic statement is equivalent to the provability in Informal Morphology$^{\mathsf{O}}$ of the counterpart of p in the sense of the translation. The reason is simple: in Informal Morphology$^{\mathsf{O}}$ it is required to *establish* a statement expressing that a refutable sentence ξ of the language of the object theory is not provable in the object theory itself. For a reasonable choice of it, in fact, there is obviously no counterpart for this sentence which is a theorem of the object theory.

The process of regimentation of our metamathematical practice carried out below aims at providing a formal system capturing the correspondence just sketched at the informal level. In order to capture this correspondence in a single formal system, a preliminary step consists in the possibility of expressing, by means of suitable definitions, syntactic and logical notions concerning the object theory O and proving basic syntactic facts concerning its language.[5] To this purpose we will still employ HF, the theory of hereditarily finite sets presented in Chapter 3. Again the reason behind the choice of HF as the *κοινή* of our syntactic reasoning, either here or in Chapter 3, is the possibility of naturally and faithfully capturing the set theoretic constructions proper of Informal Morphology$^{\mathsf{O}}$. The domain of discourse of Informal Morphology$^{\mathsf{O}}$, at least according to the picture chosen in §3.1, consists in fact in primitive symbols of $\mathcal{L}_{\mathsf{O}}$ and tuples of them. Trying to apply the reduction

[5] Such as, for instance, that a formula is different form its negation, or the conjunction of two formulas is different from both conjuncts.

procedure outlined above, we then associate these objects with elements of the class $\mathbb{G}$, that is the smallest class of closed terms obtained from 0 by self-application of the adjunction operation $\triangleleft$ and closed under pairing, defined in §3.3 (Definition 3.6). The reasons for picking out codes from $\mathbb{G}$ are also explained there. Basic theorems of HF, as well as the notion of finite sequence, readily available in HF, will do the rest.

Taken the choice of HF for granted, in the next section we will arrive at a full formalisation of the interactions between O and $\mathsf{Informal\ Morphology}^{\mathsf{O}}$ through a series of intermediate steps.

5.2 Formalising $\mathsf{Informal\ Morphology}^{\mathsf{O}}$

Besides the specific purpose represented by the fresco of informal metamathematical practice that we are aiming at, already at earlier stages of the project of formulating theories of truth with 'disentangled' syntax, the investigation of possible ways of connecting the domain of syntactic entities with the domain of discourse of the object theory was felt as a necessity:

> Although in informal metamathematics we do distinguish between syntactic and mathematical objects such as numbers and sets and the associated theories, we are usually happy to pass from the syntactic consistency statement, to its coded counterpart. [...] In the theory [...] [n.a: $\mathsf{CTD}[\mathsf{O}]$] this transition is not possible: the syntactic statements, formulated in $\mathcal{L}_{\mathsf{S}}$, are hardly related to the corresponding mathematical sentences. The reason for this is that theorising about syntactic objects is completely detached from theorising about mathematical objects. In particular, there is no obvious way to talk about (mathematical) codes of syntactic objects in [...] [n/a: $\mathsf{CTD}[\mathsf{O}]$]. To obtain a setting that is more natural than [...] [n/a: $\mathsf{CTD}[\mathsf{O}]$], one would have to add 'bridge' laws between [...] [n/a:O] and [...] [n/a: HF], axioms that allow one to connect mathematical and syntactic objects. ([Halbach 2011, p. 320])

We are now going to explore a general strategy to provide these 'bridge laws' connecting the syntactic and mathematical domain. To be sufficiently precise, however, one would need to specify the language of O. We first consider the case in which HF is also the object theory. This will enable us to formulate the arguments in quite detailed form. We the consider how to adapt ourselves to alternative choices of the object theory.

The strategy that will be shortly introduced is partly accomplished, for a slightly different setting, in an unpublished note by Jeffrey Ketland and also implicitly suggested, as we have just seen, in [Halbach 2011]. For a similar solution when the syntax theory is taken to be arithmetic and not HF, along the lines of Ketland's proposal, see [Leigh&Nicolai 2013].

5.2.1 Bridging syntax and mathematics.

We start with some notational preliminaries. Let us first introduce the two-sorted language $\mathcal{L}_{\mathsf{H{\cdot}O}}$; intuitively, $\mathcal{L}_{\mathsf{H{\cdot}O}}$ is like the $\mathcal{L}_{\mathsf{TD}}$ from Chapter 4 except for the fact that it does not feature quantification over the 'third realm' of sequences of objects from the domain of the object theory (cfr. Figure 4.1). Therefore $\mathcal{L}_{\mathsf{H{\cdot}O}}$ will contain the nonlogical constants $0, \lhd, \in$ from the alphabet of $\mathcal{L}_{\mathsf{HF}}$, nonlogical constants from the alphabet of $\mathcal{L}_{\mathsf{O}}$, which will be unspecified in this prologue, and among its logical constants, besides the usual ones, we will have—assuming a set of sorts $\mathsf{I}' = \mathsf{I} - \{sq\} = \{s, o\}$—a family of disjoint countable sets of variables

$$V' = V_s \sqcup V_o,$$

so that, as before,

$$V_o = \{v_1^o, v_2^o, \ldots\} \qquad \text{and} \qquad V_s = \{v_1^s, v_2^s, \ldots\}.$$

Consistently with the policy already adopted in the previous chapter, we denote with $i, j, k, l, m, n\ldots$, possibly with indices, variables in V_s, and with $x, y, u, v, z, \ldots$, possibly with indices, variables in V_o. We will also, abusing of notation, keep indexing nonlogical constants of $\mathcal{L}_{\mathsf{H\cdot O}}$ with elements of I'.

Given HF3, introduced in §3.3, that is the axiom schema of induction on hereditarily finite sets, we will call HF3*↾ the schema HF3 formulated in the language $\mathcal{L}_{\mathsf{H\cdot O}}$, so that arbitrary formulas of $\mathcal{L}_{\mathsf{H\cdot O}}$ are allowed to appear into the schema, although the relevant variables still belongs to V_s. The notation is inspired by the fact that HF3*↾ is actually the induction schema of CTD[O] from the previous chapter restricted to the 'syntactic' and to the 'object-theoretic' languages.

We can now define the intermediate theory H·O. For now it is not necessary to specify any particular axiomatisation of O.

Definition 5.1 (H·O)**.** H·O is the two-sorted theory formulated in the language $\mathcal{L}_{\mathsf{H\cdot O}}$. Its axioms are:

	Axioms of O
HF1	$\forall k(k = 0 \leftrightarrow \forall l(l \notin k))$
HF2	$\forall k, l, m(k = l \lhd m \leftrightarrow \forall n(n \in k \leftrightarrow n \in l \vee n = m))$
HF3*↾	$\varphi(0) \wedge \forall k \forall l(\varphi(k) \wedge \varphi(l) \rightarrow \varphi(k \lhd l)) \rightarrow \forall k \varphi(k)$
	for any $\varphi(v)$ in $\mathcal{L}_{\mathsf{H\cdot O}}$ with k, l of sort s.

H · HF^o. We aim at determining a correspondence between 'syntactic' objects, the objects named by the 'terms' of the language of Informal Morphology$^{\mathsf{O}}$ and objects that represent them in the domain of discourse of the object theory itself. Here we are concerned with the case in which the object theory O is HF itself.

For the reasons already considered in Chapter 3 and recalled in the previous section of the present, we think that HF is also a good candidate to capture an important portion of Informal Morphology$^{\mathsf{HF}}$ as it is commonly depicted. We are thus facing the situation in which on the one hand we have HF formalising a portion of the informal mathematics needed to talk about syntactic and logical properties of the system HF—we call HF^s this theory. On the other, we have HF itself and the possibility of developing in a fairly natural way some portions of the metamathematics of HF into HF itself, which will be called, from now on, HF^o.

Our purpose is then to construct a theory, containing the 'syntactic' HF—that will be labelled HF^s—and the 'mathematical' HF—denoted with HF^o—which is capable of formalising the two-level correspondence between (i) objects of Informal Morphology$^{\mathsf{HF}}$ and finite sets, (ii) predicates of Informal Morphology$^{\mathsf{HF}}$ and a formulas of HF sketched above. In practice, we create an extension of $\mathsf{H{\cdot}HF}^o$ and define, within the resulting theory, a relative interpretation of HF into HF itself. We might in fact describe our method as an attempt to formalise the process of relatively interpreting a variant of HF^o in HF^o itself according to the well-known definition of relative interpretation given in [Tarski et alii 1953]. As we said already, given the current choice of the syntax theory and the object theory, the definition of this interpretation will turn out to be fairly trivial. However, the work carried out here may be useful to deal with less trivial cases.

As starting point of our formalisation, we then have $\mathsf{H{\cdot}HF}^\circ$ formulated in the language $\mathcal{L}_{\mathsf{H\cdot HF}^\circ}$. To relate the two domains of quantifiers of $\mathcal{L}_{\mathsf{H\cdot HF}^\circ}$, we expand $\mathcal{L}_{\mathsf{H\cdot HF}^\circ}$ with a binary predicate symbol B^{hf} of sort (s,o), and extend $\mathsf{H{\cdot}HF}^o$ with suitable axioms governing its behaviour. As Ketland emphasised,[6] the strategy is reminiscent of the reduction of scientific theories containing mathematical vocabulary to their nominalistically acceptable counterparts, presented for instance in [Burgess&Rosen 1997, §I.B.2]. In order to generate a correspondence between the two domains, a natural guess is to relate entities belonging to the domain of one copy of HF to the other. Therefore we first suggest to extend $\mathsf{H{\cdot}HF}^o$ formulated in $\mathcal{L}_{\mathsf{H\cdot HF}^\circ} \cup \{B^{\mathsf{hf}}\}$ with the axioms:

$$\mathsf{B1}^{\mathsf{hf}} \qquad \forall x\big(B^{\mathsf{hf}}(x,\mathrm{o}^s) \leftrightarrow\; x = \mathrm{o}^o\big)$$

$$\mathsf{B2}^{\mathsf{hf}} \qquad \forall x \forall k,l\big(B^{\mathsf{hf}}(x, k \lhd^s l) \leftrightarrow \exists y,z\big(B^{\mathsf{hf}}(y,k) \wedge B^{\mathsf{hf}}(z,l) \wedge x = y \lhd^o z\big)\big)$$

[6]In private communication.

Let us call the resulting theory $\mathsf{H{\cdot}HF}^{+}$. Its axioms then are, besides $\mathsf{B1}^{\mathsf{hf}}$ and $\mathsf{B1}^{\mathsf{hf}}$, the axioms of the 'syntactic' theory HF^{s}, the axioms of the 'mathematical' theory HF^{o}, and the induction schema $\mathsf{HF3}^{*}\!\restriction$ extended to allow occurrences of the new relation symbol B^{hf} in its instances. Let us call this new induction schema $\mathsf{HF3}^{*}\!\restriction^{B}$. We have

Lemma 5.1. $\mathsf{H{\cdot}HF}^{+} \vdash \forall k \exists x \big[B^{\mathsf{hf}}(x,k) \wedge \forall y (B^{\mathsf{hf}}(y,k) \rightarrow x = y) \big]$.

Proof. The proof is obtained by applying $\mathsf{HF3}^{*}\!\restriction$ to the formula above. *qed.*

Lemma 5.1 tells us that any object k in the sense of a model of HF^{s} has a unique counterpart in the universe described by HF^{o}. In other words, $\mathsf{H{\cdot}HF}^{+}$ defines an injection of 'syntactic' objects into 'mathematical' objects.

One step towards the formalisation of the correspondence between objects of Informal Morphology$^{\mathsf{HF}^{o}}$ and their counterparts seems to have been accomplished. However, something is still missing. Let us consider in fact the structure

$$\mathcal{D}' = \{\mathbb{H}, |\mathcal{M}|, \mathcal{I}^{o} \cup \mathcal{I}^{s} \cup \mathcal{I}^{B}\}$$

where $\mathbb{H}$ is the domain standard model of HF^{s}, the syntactic HF, $|\mathcal{M}|$ is the domain of an ω-nonstandard model $\mathcal{M}$ of HF^{o}, the object theory, and $\mathcal{I}^{B}$ is the interpretation of the new symbol B^{hf}, that is an injection $f\colon \mathbb{H} \rightarrow |\mathcal{M}|$. Let us recall what we aim at: a formula $\varphi^{*}(x)$ of the language of HF^{o} which translates a formula $\varphi(k)$ of the language of HF^{s}, with $B^{\mathsf{hf}}(x,k)$, has to be true in the model $\mathcal{D}'$ exactly of the entities which are the counterparts of the objects in the domain of discourse of HF^{s} which satisfy the formula $\varphi(k)$ of $\mathcal{L}_{\mathsf{HF}^{s}}$, that is of x. Now obviously in $\mathsf{H{\cdot}HF}^{+}$ we can carry out a translation of the language of HF^{s} into the language of HF^{o}, so to associate to any formula $\varphi(k)$ as above a counterpart $\varphi^{*}(x)$. However, if $\varphi^{*}(x)$ is such that $\mathcal{D}' \vDash \varphi(\overline{\alpha})$ for infinitely many ordinals $\alpha \in |\mathcal{M}|$, by the Overspill Principle form the previous chapter we notice that $\mathcal{D}'$ sees $\varphi^{*}(x)$ true of more than just their counterparts in $\mathbb{H}$ according to the defined injection; in particular we will have $\mathcal{D} \vDash \varphi^{*}(\overline{c})$ for some nonstandard $c \in |\mathcal{M}|$, therefore by Lemma 5.1 there will be objects in $|\mathcal{M}|$ satisfying $\varphi^*(x)$ which do not have a counterpart in $\mathbb{H}$. Now if one applies this observation to our framework, we notice that there might be objects in $|\mathcal{M}|$ so that, for instance $\mathcal{D}'$ recognises as codes of sentences^o of $\mathcal{L}_{\mathsf{O}}$—that is we may have $\mathcal{D}' \vDash \mathsf{Sent}^o_{\mathcal{L}_{\mathsf{O}}}(\overline{a})$—which do not have twins in $\mathbb{H}$—that is there is no m in $\mathbb{H}$ such that the pair (a, m) is in the extension of B^{hf}. This can be disastrous for the purpose of establishing a faithful correspondence between statements of syntactic nature belonging to the two languages.

In order to faithfully capture the development of the syntax for HF^o carried out in HF^s, the injection induced by $B^{\mathsf{hf}}(x, k)$ is not sufficient. We need to supplement $\mathsf{H{\cdot}HF}^+$ with further axioms which would determine a bijection between the 'syntactic' and the 'mathematical' universe. To this purpose, it would suffice to stipulate that any object in the range of quantifiers of HF^o has a unique counterpart in the range of quantifiers of HF^s: in particular, we may assume that for any x of sort o there is a unique k of sort s such that $B^{\mathsf{hf}}(x, k)$.

Let us thus define, the extension of $\mathsf{H{\cdot}HF}^+$, which we will call $\mathsf{H{\cdot}HF}^B$, obtained by supplementing its axioms with the clause just discussed.

Definition 5.2 ($\mathsf{H{\cdot}HF}^B$)**.** $\mathsf{H{\cdot}HF}^B$ is the theory formulated in $\mathcal{L}_{\mathsf{H{\cdot}HF}^o}$ whose axioms are:

$\mathsf{B1}^{\mathsf{hf}}$ $\quad \forall x\big(B^{\mathsf{hf}}(x, \mathsf{o}^s) \leftrightarrow x = \mathsf{o}^o\big)$

$\mathsf{B2}^{\mathsf{hf}}$ $\quad \forall x \forall k, l\big(B^{\mathsf{hf}}(x, k \lhd^s l) \leftrightarrow \exists y, z\big(B^{\mathsf{hf}}(y, k) \wedge B^{\mathsf{hf}}(z, l) \wedge x = y \lhd^o z\big)\big)$

$\mathsf{B3}^{\mathsf{hf}}$ $\quad \forall x \exists k\big[B^{\mathsf{hf}}(x, k) \wedge \forall l(B^{\mathsf{hf}}(x, l) \to k = l\big]$

$\mathsf{B3}^{\mathsf{hf}}$ can be considered the dual of the statement of Lemma 5.1. It states that any object in the mathematical domain has a unique counterpart in the realm of syntactic objects.

$\mathsf{H{\cdot}HF}^B$, given the bijection determined by its axioms, is able to establish the connection between syntactic predicates and a a subclass of formulas of the object theory considered in point (ii) of our informal characterisation above.

We start with a useful lemma,

Lemma 5.2.

(i) $B^{\mathsf{hf}}(x,k) \wedge B^{\mathsf{hf}}(y,l) \rightarrow (x = y \leftrightarrow k = l)$

(ii) $B^{\mathsf{hf}}(x,k) \wedge B^{\mathsf{hf}}(y,l) \rightarrow (x \in^{o} y \leftrightarrow k \in^{s} l)$.

Proof. (i) follows from a straightforward induction. (ii) is obtained by applying $\mathsf{HF3}^{*}\restriction$ to

$$\varphi(l) :\leftrightarrow B^{\mathsf{hf}}(x,k) \wedge B^{\mathsf{hf}}(y,l) \rightarrow (x \in^{o} y \leftrightarrow k \in^{s} l),$$

for arbitrary k, x, y. *qed.*

Proposition 5.1. For any formula $F(k_1, ..., k_n)$ of $\mathcal{L}_{\mathsf{H}\cdot\mathsf{HF}^o}$ in which only variables in V_s (free or bound) and nonlogical constants from the vocabulary of HF^s can occur, there is a formula $F^{*}(x_1, ..., x_n)$ of $\mathcal{L}_{\mathsf{H}\cdot\mathsf{HF}^o}$ with only (free or bound) variables in V_o and nonlogical constants from the alphabet of the language of HF^o such that $\mathsf{H}\cdot\mathsf{HF}^B$ proves

$$\forall x_1, ..., x_n\big(F^{*}(x_1, ..., x_n) \leftrightarrow \exists k_1, ..., k_n(B^{\mathsf{hf}}(x_1, k_1) \wedge .... \wedge B^{\mathsf{hf}}(x_n, k_n) \wedge F(k_1, ..., k_n))\big).$$

Proof. By induction on the complexity of the syntactic formula $F(k_1, ..., k_n)$.

For the case in which $F(k_1, ..., k_n)$ is an atomic formula, we have the following cases. If $F(k_1, ..., k_n)$ is $k_1 = 0^s$, then the claim is equivalent to $\mathsf{B1}^{\mathsf{hf}}$. If $F(k_1, ..., k_n)$ is $k_1 = k_2 \lhd^{s} k_3$, the left-to-right direction is straightforward (in particular, if k_1 is different from $k_2 \lhd^{s} k_3$ for all k_2, k_3, then $k_1 = 0^s$). The right-to-left direction is obtained by employing Lemma 5.2. The case in which $F(k_1, ..., k_n)$ is $k_1 \in^{s} k_2$ follows from Lemma 5.2. The cases in which $F(k_1, ..., k_n)$ is a complex formula are straightforward and we skip their proofs. *qed.*

As an immediate consequence of Proposition 5.1, we have that

Corollary 5.1. For any sentence A containing only quantification over variables in V_s and non logical vocabulary from the alphabet of the language of HF^s, there is a sentence A^{*}

obtained according to the translation induced by Proposition 5.1 such that

$$\mathsf{H{\cdot}HF}^{B} \vdash A \leftrightarrow A^{*}.$$

A fortiori, we will have, by Σ_1-completeness (Proposition 3.3):

Corollary 5.2. For any Σ_1-sentence A of the language of HF^{s} true in the standard model $\mathcal{H}$ of HF, we have

$$\mathsf{H{\cdot}HF}^{B} \vdash A^{*}.$$

Alternative choices of O. If we consider different choices of the object theory O and of its language, the strategy outlined above can be replicated with suitable modifications.

An obvious shortcut would be to calibrate the choice of theory formalising Informal Morphology$^{\mathsf{O}}$ to the language of O. If O is, for instance, Peano Arithmetic, we might consider Elementary Arithmetic (see [Beklemishev 2005, §1]), or $I\Sigma_1$ to formalise Informal Morphology$^{\mathsf{PA}}$. We would then construct the two-sorted theory $I\Sigma_1 \cdot \mathsf{PA}$ along the lines of $\mathsf{H{\cdot}O}$ in Definition 5.1, expand its language with a binary predicate symbol of sort (s, o), such as B^{hf} above, and axiomatise it to determine a mapping of the 'syntactic' nonlogical constants of the 'syntactic' language, that is the language of $I\Sigma_1$ in the case at issue, into the ones belonging to the 'mathematical' language, PA in this case, so that the results obtained in the previous subsection hold. Similarly, if O is ZF, we may formalise Informal Morphology$^{\mathsf{ZF}}$ in $\mathsf{KP}\omega$—to mention a theory which nicely develops the syntax for ZF—and proceed accordingly.

On the other hand, if we consider HF as the obvious choice for the formalisation of Informal Syntax$^{\mathsf{O}}$, holding that it is better to stick with it when it is possible, we can still carry out the strategy outlined above, at least in the cases in which O is formulated in the language of arithmetic or in the language of set theory.

For the latter, for instance, in which the only nonlogical constant will be $\in^{o}$, let us consider for instance the theory $\mathsf{H{\cdot}Z}$ constructed according to Definition 5.2 where, for simplicity, Z is a pure extension of Zermelo Fraenkel set theory minus the axiom of infinity ($\mathsf{ZF} - \mathsf{Inf}$). This choice of O eliminates possible doubts concerning the possibility of mimicking the development of the syntax of O in O itself. If one requires Z to be finitely axiomatised,[7] one has to make sure that the formalisation of syntactic and logical notions concerning Z can be carried out in it in a 'natural' way. After expanding the two-sorted language $\mathcal{L}_{\mathsf{H{\cdot}Z}}$—which will maintain the abbreviations for variables of each sort already introduced for $\mathsf{H{\cdot}HF}$—with the symbol B^{z} of arity 2 and of sort (s, o), we can define the theory $\mathsf{H{\cdot}Z}^B$ by adding to the set of axioms of $\mathsf{H{\cdot}Z}$ the axioms

$\mathsf{B1}^z$ $\quad B^z(x, 0^s) \leftrightarrow (\forall y \in x)(y \in y)$;

$\mathsf{B2}^z$ $\quad B^z(x, k \lhd^s l) \leftrightarrow \exists yz(B^z(y, k) \wedge B^z(z, l) \wedge (\forall u \in x)(u \in y \vee u = z))$;

$\mathsf{B3}^z$ $\quad \exists! k\, B^z(x, k)$.

As before, in the presence of a sufficiently strong O, $\mathsf{B3}^z$ is needed to rule out possible nonstandard interpretations of the syntax of Z developed in Z itself. It is immediate to see that the development of $\mathsf{Informal\ Morphology}^{\mathsf{z}}$ in HF can be replicated in Z, so that we can prove results analogous to the ones presented above.

The case in which O is formulated in the language $\mathcal{L}_{\mathsf{A}} = \{0^o, S^o, +^o, \times^o\}$ of arithmetic can be treated in a similar way. Again to avoid unnecessary complications we consider the case in which O is PA—which will be denoted with PA^o. The new binary predicate B^{pa} of sort (s, o) would be axiomatised following the idea of mimicking the Ackermann interpretation (cfr. §3.2.1); therefore the first axiom will guarantee that to the constant 0^s for the empty set of HF^s is assigned the 'mathematical' 0^o, that is

$\mathsf{B1}^{\mathsf{pa}}$ $\quad B^{\mathsf{pa}}(x, 0^s) \leftrightarrow x = 0^o$.

[7] As Ketland does in his unpublished note.

In order to formulate the next axiom we recall the the functional expression of the language of arithmetic $\mathsf{bit}(x,y)$ denotes the unique $v \leq 1$ such that

$$(\exists u \leq y)(\exists w < 2^x)(y = 2^{x+1} \cdot u + 2^x \cdot v + w).$$

and that the Ackermann interpretation assigns to $x \in y$ in the language of HF the formula $\mathsf{bit}(x,y) = 1$, expressing that x occurs in the binary representation—which is unique for any natural number—of y. Therefore we have the axiom

$$\mathsf{B2}^{\mathsf{pa}} \qquad B^{\mathsf{pa}}(x, k \lhd^s l) \leftrightarrow \exists y, z(B^{\mathsf{pa}}(y,k) \wedge B^{\mathsf{pa}}(z,l) \wedge \forall u(\mathsf{bit}(u,x) = 1 \leftrightarrow (\mathsf{bit}(u,y) = 1 \vee u = z))).$$

The axiom states that the number x represents the finite set $k \lhd^s l$ if and only if there are numbers y and z which represent the finite sets k, l respectively, and that a number u occurs in the binary representation of x if and only if either it occurs in the binary representation of y or it is equal to z. Summarising, the axioms of the theory $\mathsf{H{\cdot}PA}^B$ would be:

$\mathsf{B1}^{\mathsf{pa}}$ $\quad B^{\mathsf{pa}}(x, 0^s) \leftrightarrow x = 0^o$.

$\mathsf{B2}^{\mathsf{pa}}$ $\quad B^{\mathsf{pa}}(x, k \lhd^s l) \leftrightarrow \exists yz(B^{\mathsf{pa}}(y,k) \wedge B^{\mathsf{pa}}(z,l) \wedge \forall u(\mathsf{bit}(u,x) = 1 \leftrightarrow (\mathsf{bit}(u,y) = 1 \vee u = z)))$.

$\mathsf{B3}^z$ $\quad \exists! k\, B^z(x,k)$.

We are then be able to prove the analogues of Proposition 5.1 and Corollaries 5.1 and 5.2.

5.3 Formalising Informal Metamathematical Practice.

If we establish that a formula φ of $\mathcal{L}_{\mathsf{O}}$ is a theorem of the mathematical theory O, in our informal metamathematical discussion we can conceive this claim as a syntactic relation occurring between the axioms of O, its rules of inference and the formula φ; this relation is usually denoted with $\mathsf{O} \vdash \varphi$. According to our description,[8] '$\vdash$' has to be understood as an expression pertaining to the mathematician's English, thus to Informal Morphology$^{\mathsf{O}}$, expressing in our particular case the existence of a derivation, a finite sequence of formulas, from the axioms of O of a formula φ of $\mathcal{L}_{\mathsf{O}}$. When Informal Morphology$^{\mathsf{O}}$ is replicated within a sufficiently strong O, then we have seen that it is possible, under fairly general assumptions, to construct a formula $\mathsf{Bew}^{o}_{\mathsf{O}}(x)$, expressing within O the relation expressed by $\mathsf{O} \vdash \varphi$ in the language of Informal Morphology$^{\mathsf{O}}$. In particular, it is usually required that the following holds:

$$\mathsf{O} \vdash \varphi \quad \text{if and only if} \quad \mathsf{O} \vdash \mathsf{Bew}^{o}_{\mathsf{O}}(\overline{\varphi}) \tag{5.1}$$

The correspondence is faithfully captured by Proposition 5.1.

For instance, in the case of $\mathsf{H{\cdot}HF}^{B}$, a claim along the lines of the lefthand side of (5.1), belonging to Informal Morphology$^{\mathsf{HF}}$, is formalised by $\mathsf{Bew}^{s}_{\mathsf{HF}^{o}}(\ulcorner \sigma \urcorner)$. Proposition 5.1 thus yields, provably within $\mathsf{H{\cdot}HF}^{B}$,

$$\mathsf{Bew}^{s}_{\mathsf{HF}^{o}}(\ulcorner \varphi \urcorner) \leftrightarrow (\mathsf{Bew}^{s}_{\mathsf{HF}^{o}}(\ulcorner \varphi \urcorner))^{*} \tag{5.2}$$

We remains to be shown is that the sentence $(\mathsf{Bew}^{s}_{\mathsf{HF}^{o}}(\ulcorner \varphi \urcorner))^{*}$ can be considered a canonical provability predicate of HF^{o}. One may require to show that

$$\mathsf{HF}^{o} \vdash \varphi \Leftrightarrow \mathsf{HF}^{o} \vdash (\mathsf{Bew}^{s}_{\mathsf{HF}}(\ulcorner \varphi \urcorner))^{*}. \tag{5.3}$$

(5.3), however, holds by virtue of a more general observation, again a corollary of Proposition 5.1, according to which for any formula $A(k_1, \dots, k_n)$ expressing a canonical syntactic relation among finite sets coding expressions of $\mathcal{L}_{\mathsf{HF}^{o}}$, the formula $A^{*}(x_1, \dots, x_n)$ which is equivalent

[8] Cfr. also [Boolos 1993, xiiiff].

to

$$\exists k_1, \dots, k_n (B^{\mathsf{hf}}(x_1, k_1) \wedge \dots \wedge B^{\mathsf{hf}}(x_n, k_n) \wedge A(k_1, \dots, k_n)),$$

expresses the same syntactic relation constructed by mimicking the construction of $A(k_1, \dots, k_n)$ in HF^s. For instance, to the formula $\mathsf{Bew}^s_{\mathcal{L}_{\mathsf{HF}^o}}(k)$ of $\mathcal{L}_{\mathsf{HF}^s}$ expressing that k codes a proof in HF^o, by Proposition 5.1, we associate a formula $\psi(x)$ of $\mathcal{L}_{\mathsf{HF}^o}$ such that

$$\mathsf{H{\cdot}HF}^B \vdash \forall x(\psi(x) \leftrightarrow \exists k(B^{\mathsf{hf}}(x, k) \wedge \mathsf{Bew}^s_{\mathcal{L}_{\mathsf{HF}^o}}(k))). \tag{5.4}$$

It can then be shown, by employing (5.3), that

$$\mathsf{HF}^o \vdash \varphi \Leftrightarrow \mathsf{HF}^o \vdash \psi(\overline{\varphi}). \tag{5.5}$$

To employ a suggestive notation, we might call $\mathsf{Bew}^o_{\mathsf{HF}^o}(x)$ this $\varphi(x)$. There is no need to emphasise that $\mathsf{Bew}^o_{\mathsf{HF}^o}(x)$ so constructed actually mirrors the construction of $\mathsf{Bew}^s_{\mathsf{HF}^o}(k)$, that is it is true of the entities which are the representations, in the sense of $\mathsf{H{\cdot}HF}^B$, of the entities of which $\mathsf{Bew}^s_{\mathsf{HF}^o}(k)$ holds.

In dealing with object theories formulated in other languages, although an analog of Proposition 5.1 would still hold, we would need to ensure that the formalisation of Informal Morphology$^{\mathsf{O}}$ carried out in HF^s can be faithfully preserved. It is not possible to make specific claims without referring to a particular axiomatisation of the object theory; at any rate, what is required is that the syntax of O, as developed in HF^s, can be replicated in O, in the sense that we can set up a bijection between the domain of discourse of HF^s and a subset of the domain of discourse of O, and a correspondence between each 'syntactic' formula of HF^s concerning O and a formula φ of $\mathcal{L}_{\mathsf{O}}$ so that φ satisfies the conditions stated above.

$\mathsf{CTD[O]}^B$. The motivation behind the construction of $\mathsf{H{\cdot}O}^B$ was the possibility of constructing a formal system which was able to mirror the interaction between informal and formal aspects of our metamathematical reflection over the object theory O. In particular, we wanted to provide a formalisation of what Halbach calls 'notational simplification' (see quotation above), that is the process of identifying expressions with sets or numbers carried out when from informal metamathematical discussion we move to its formalisation. We have seen in fact that theories constructed along the lines of $\mathsf{H{\cdot}O}^B$ are able to link informal statements of syntactic nature, now formalised in HF^s, and their object theoretic counterparts.

For instance, if we consider a theorem φ of HF^o, we will have that $\mathsf{Bew}^s(\ulcorner\varphi\urcorner)$ is provable in HF^s by Proposition 3.9—that is what we called 'Provability Condition' in Ch. 3. Therefore $\mathsf{Bew}^o_{\mathsf{HF}^o}(\overline{\varphi})$, constructed according to the translation determined by the axioms for B^{hf}, is a consequence of $\mathsf{H{\cdot}HF}^B$ by Proposition 5.1. Similarly, φ is a formula of $\mathcal{L}_{\mathsf{HF}}$, then $\mathsf{Fml}^s_{\mathcal{L}_{\mathsf{HF}^o}}(\ulcorner\varphi\urcorner)$ is true in the standard model $\mathcal{H}$ of HF^s, and thus provable in HF^s by Proposition 3.3 (Σ_1-completeness, cfr. Ch. 3). Proposition 5.1 then entails that $\mathsf{Fml}^o_{\mathcal{L}_{\mathsf{HF}^o}}(\overline{\varphi})$ is provable in $\mathsf{H{\cdot}HF}^B$. The same applies to other Σ_1-formulas of $\mathcal{L}_{\mathsf{HF}^s}$ expressing syntactic relations among expressions of $\mathcal{L}_{\mathsf{HF}^o}$.

It seems thus, more in general, that a good deal of $\mathsf{Informal\ Morphology}^{\mathsf{O}}$, as formalised in HF^s, can be recovered in $\mathsf{H{\cdot}HF}^B$.

A noticeable component of our informal metamathematical reflection *on* O, however, is represented by the belief in the consistency of O. In fact:

> Although in informal metamathematics we do distinguish between syntactic and mathematical objects such as numbers and sets and the associated theories, we are usually happy to pass from the syntactic consistency statement, to its coded counterpart. For instance, it is uncontroversial that if ZF is consistent, then the consistency statement $\mathsf{Con}^o_{\mathsf{ZF}}$ [...] is also true. [Halbach 2011, p. 320]

Translated in our current idiom, this means that the (informal) consistency statement for O ought to be a *theorem* of our formalisation of the metamathematical reasoning on O; as a result, an adequate formalisation of $\mathsf{Informal\ Morphology}^{\mathsf{O}}$ would need to determine a theory in which this syntactic consistency statement is provable.

To complete the process of formalisation, we thus propose a theory which (i) contains O as a subtheory; (ii) formalises the informal characterisation of the syntax of O; (iii) establishes our belief in the soundness of O; (iv) associates formal and informal claims in a suitable way. If conditions (i) and (ii), and (iv) are fulfilled by $\mathsf{H{\cdot}O}^B$ already, condition (iii) cannot be satisfied, for a reasonable choice of O. To establish this, (simplifications of) the arguments along the lines of the ones employed in Theorems 4.1 and 4.2 from Chapter 4, namely a well-behaved relative interpretation of $\mathsf{H{\cdot}HF}^B$ in O can be employed.

A natural candidate to capture condition (i)-(iv) is represented by the theory $\mathsf{CTD[O]}^B$. It can be considered as an extension of $\mathsf{CTD[O]}$, introduced in Definition 4.3, which is obtained by expanding its (three-sorted) language $\mathcal{L}_{\mathsf{TD}}$ with the binary predicate B° of sort (s, o), allowed to appear into the induction schema $\mathsf{HF3}^*$, and suitable axioms devised following the method sketched above. Equivalently, from a different point of view, $\mathsf{CTD[O]}^B$ can be considered as an extension of $\mathsf{H{\cdot}O}^B$ obtained by expanding its language with a new set of variables labeled by the new sort sq, function symbols $.(.)$ and D of type $(sq, s) \to o$ and a binary satisfaction predicate Sat of sort (sq, s) axiomatised along the lines of Table 4.1, and allowing the new vocabulary to appear into the induction schema $\mathsf{HF3}^* \restriction$.

As we have seen, the theory $\mathsf{CTD[O]}$ introduced in the previous chapter determines a somewhat puzzling splitting phenomenon concerning consistency statements for O, so that we can fix a syntactic consistency statement $\mathsf{Con}^s_{\mathsf{O}}$ formalised in the language of HF and a (canonical) mathematical counterpart $\mathsf{Con}^o_{\mathsf{O}}$. The conservativity of $\mathsf{CTD[O]}$ and $\mathsf{CTD[O]}{+}\mathsf{AxT}^s_{\mathsf{O}}$ over O (Cfr. Theorem 4.3 and Corollary 4.6), where $\mathsf{AxT}^s_{\mathsf{O}}$ is the $\mathcal{L}_{\mathsf{TD}}$-sentence expressing that all the axioms of O are true by means of a $\mathcal{L}_{\mathsf{HF}}$ representation of the axioms of O, is an undisputed sign that the two consistency assertions are in all respects unrelated. On the other hand, it appears to be clear, as witnessed by Halbach's quotation above, that in metamathematical practice the informal statement expressing that no contradiction can be derived from the axioms of O is somewhat connected with its formalisation in the language of O.

Let us concentrate on the comfortable case of $\mathsf{CTD[HF]}^B$. Corollary 4.5 tells us that $\mathsf{CTD[HF]}^B$+$\mathsf{AxT}^s_{\mathsf{O}}$ proves $\mathsf{Con}^s_{\mathsf{HF}}$, that is the consistency statement for the 'mathematical' HF^o formulated in the language of HF^s. By Proposition 5.1, we have that $(\mathsf{Con}^s_{\mathsf{HF}})^*$, the $\mathcal{L}_{\mathsf{HF}^o}$-sentence induced by the 'coding axioms', is an intensionally correct consistency statement for HF^o—and that it is not provable in HF^o itself by Theorem 3.2. Following our notational conventions, we denote $(\mathsf{Con}^s_{\mathsf{HF}^o})^*$ with the more suggestive $\mathsf{Con}^o_{\mathsf{HF}^o}$. Does $\mathsf{CTD[HF]}^B$ capture conditions (i)-(iv)? Evidently not as, although we can prove in it the equivalence between different consistency statements, we cannot prove $\mathsf{Con}^s_{\mathsf{HF}}$ in it, by suitable modifications of Theorem 4.2.

By Corollary 5.1, if O is a finitely axiomatised fragment of HF, or a finitely axiomatised portion of set theory in which the development of the syntax for O carried out in HF^s can be mimicked, then $\mathsf{CTD[O]}^B$ can be taken to be a sufficient, for our purposes, formalisation of informal metamathematical reflection concerning O. In fact, in addition to the nice features displayed by $\mathsf{H{\cdot}O}^B$, we can have

Corollary 5.3. Let O be a sufficiently strong finite fragment of HF, then

$$\mathsf{CTD[O]}^B \vdash \mathsf{Con}^o_{\mathsf{O}^o}.$$

If O is instead schematically axiomatised, for instance HF itself—but similar remarks hold clearly for PA and ZF—we need instead to focus on the theory $\mathsf{CTD[HF}^o]^B + \mathsf{AxT}^s_{\mathsf{HF}^o}$. In fact, Corollary 4.8 tells us that $\mathsf{Con}^s_{\mathsf{HF}^o}$ is a theorem of $\mathsf{CTD[HF}^o]^B + \mathsf{AxT}^s_{\mathsf{HF}^o}$; by Corollary 5.1 then,

Corollary 5.4. $\mathsf{CTD[HF}^o]^B + \mathsf{AxT}^s_{\mathsf{HF}^o} \vdash \mathsf{Con}^o_{\mathsf{HF}^o}$.

In conclusion, if on the one hand $\mathsf{CTD[HF]}^B + \mathsf{AxT}^s_{\mathsf{HF}}$ displays the interesting property of formalising the usual interactions between informal and formal metamathematical claims surrounding the object theory, HF^o in this case, on the other it loses another noticeable property possessed by theories of truth constructed with a 'disentangled' syntax:

Corollary 5.5. $\mathsf{CTD}[\mathsf{HF}]^B + \mathsf{AxT}^s_{\mathsf{HF}}$ is not conservative over HF^o.

In the previous chapter we have considered one possible route to take if one wanted to strengthen theories such as $\mathsf{CTD}[\mathsf{PA}]$, $\mathsf{CTD}[\mathsf{ZF}]$ or $\mathsf{CTD}[\mathsf{HF}]$: we extended schemata of the object theory by allowing instances containing arbitrary formulas of $\mathcal{L}_{\mathsf{TD}}$. Theorem 4.4 then told us that the resulting theories became proof-theoretically stronger than the initial theories. The strategy behind Theorem 4.4, however, works only if the object theory is schematically axiomatised: in the case of finitely axiomatised object theories, in fact, simply there are no schemata to be expanded. The strengthening of $\mathsf{CTD}[\mathsf{O}]$ resulting from the addition of coding axioms introduced above represents a method to transform a conservative theory of truth with disentangled syntax into a nonconservative one (over the object theory O) which applies also to the case in which the object theory O is finitely axiomatised.

As we have seen, however, there is a trade-off: the strategy of adding coding axioms to theories of truth with disentangled syntax over schematically axiomatised object theories leads to a nonconservative extension of the object theory only if we focus on the theory $\mathsf{CTD}[\mathsf{O}]+\mathsf{AxT}^s_{\mathsf{O}}$ where the claim 'all axioms of O are true'—employing a suitable representation of the notion of being an axiom of O in the language of HF^s—has been added to the theory of truth, whereas in theories such as $\mathsf{CTD}[\mathsf{O}]^+$ this syntactic claim followed from the extension of the schemata of the object theory.

The role of the statement $\mathsf{AxT}^s_{\mathsf{O}}$, together with further philosophical implications of the theories considered above, will be discussed in the next chapter.

6 The Trace of the True

In the present chapter we try to provide a philosophical assessment of our work on theories of truth with 'disentangled' syntax. We first give a fuzzy description of the deflationary standpoint about truth. We then review the recent debate about the so-called 'conservativeness argument', which tries to reject the deflationary viewpoint by matching the explanatory power of the truth predicate with the deductive strength of formal theories of truth. In the concluding part, we try to clarify some aspects of the discussion by invoking the theories introduced in earlier parts of this work.

6.1 Deflationism.

Deflationism can be roughly characterised as a philosophical standpoint on the nature of truth based upon the denial of the existence of a *single* property shared by truths belonging to different areas of scientific enquiry. This unofficial gloss is vague enough to generate much confusion.

One immediate worry related to our rough understanding of deflationism lies in a possible overlap with a form of *pluralism*.[1] A pluralist would in fact also deny the existence of a unique property setting the necessary and sufficient conditions for a sentence to be true, although she will hold that such properties exist only relative to different areas of discourse. The deflationist, as we shall see shortly, cannot agree with the latter claim.

[1]For an up to date and a detailed exposition of the current debate on pluralist theories of truth, see [Pedersen&Wright 2013a]. For an introductory survey, we refer to [Pedersen&Wright 2013b].

A slightly better approximation to a general characterisation of the deflationary standpoint about truth forces one to focus on the T-schema

$$T\ \overline{\sigma} \text{ if and only if } \sigma \tag{6.1}$$

where σ is a metavariable for sentences and $\bar{\cdot}$ a name forming device.[2] According to some versions of deflationism, the T-schema fixes the meaning of the truth predicate and represents an exhaustive account of the nature of truth.

We do not aim at philological accuracy. From now on, when referring to the fictional character named 'deflationist', we will refer to a philosopher who agrees on the following hallmarks:

(i) Truth is not a genuine property.
(ii) The truth predicate is a logico-linguistic device for disquotation and a *non dispensable* tool for expressing indirect endorsement and generalisations.
(iii) The notion of truth does not play any role in scientific and philosophical explanations beyond the ones expressed by (i),(ii).

In the rest of the present section, we briefly motivate the choice of (i)-(iii).

The doctrine that truth is no property dates back to the early days of analytic philosophy. In the most straightforward sense, we might read (i) as stating that truth is not even a predicate in the syntactic sense. This understanding of (i) has in turn two possible interpretations: one the one hand we can focus on subject-predicate syntactic constructions of English; on the other, we can focus on the formalisation of 'is true' in predicate logic. The first sub-interpretation seems untenable: even theories of truth that propose a revision of current English, in which the truth predicate is interpreted as a 'proform' or an operator,

[2] For coherence with the rest of the work, we formulate the T-schema employing sentences, although some diquotationalists, for instance [Horwich 1998], consider formulations involving propositions.

have equivalent formulations in terms of subject-predicate constructions. The second sub-interpretation, on the other hand, has been defended by many authors.

[Frege 1892], [Williams 1976] and [Grover, et alii 1975], for different reasons, all support this claim. If, according to Frege, the truth-value is an object and cannot be part of a 'thought',[3] Williams, building on [Prior 1971], formalises 'is true' in terms of substitutional quantification. Along similar lines, [Grover, et alii 1975] propose a refinement of Ramsey's understanding of 'is true' by means of second-order propositional quantifiers.[4] Their *prosentential theory of truth* justifies the use of this special sort of quantification on the grounds of a more general characterisation of truth as a linguistic device that simulates the anaphoric use of pronouns. Strawson in [Strawson 1949] also holds the view that truth does not apply to sentences or to any other object; although he does not consider a formal rendering of 'is true', it might be plausible to look at substitutional quantification as a natural tool to employ in a formalisation of Strawson's first proposal.

In a milder reading, (i) states that although truth can be part of a subject-predicate construction, it is no predicate in a *semantical* sense as there are no objects to which it can be ascribed.[5] This second rendering of (i) is also fairly cautious as it does not involve any reference to properties: the 'special' status of the truth predicate in fact, can be appreciated from a point of view which is neutral with respect to philosophical attitudes concerning the existence of properties. Since many contemporary deflationists have endorsed this less radical interpretation of (i), from now on we will employ the expression 'the truth predicate' with a *syntactic*, innocuous reading in mind.

The truth predicate has in fact been considered not to be a genuine predicate in a semantical sense as it reacts in a distinctive way to translations. This feature of the truth predicate has been emphasised already by [Strawson 1949] and it plays an important role in modern

[3]Cfr. [Frege 1892]: in particular, a *thought* is 'not the subjective performance of thinking but its objective content, which is capable of being the common property of several thinkers'.

[4]For Ramsey's proposal, see [Ramsey 1927].

[5]This second reading of 'is true' is called *semantical* in [Halbach 2001a].

disquotationalism. Let us consider the English sentence

(6.2) 'Lilies are flowers' is true if and only if lilies are flowers,

If 'is true' were applied to specific strings of symbols—sentences in particular—then the translation of (6.2) into Italian would be

(6.3) 'Lilies are flowers' é vero se e solo se i gigli sono fiori

However, according to Strawson, a correct translation of (6.2) would instead be

(6.4) 'I gigli sono fiori' é vero se e solo se i gigli sono fiori.

Thus 'is true' would react in a quite different way to translations if compared to predicates which genuinely apply to the English sentence 'Lilies are flowers' such as 'is grammatical English', for instance. Strawson does not specify the reasons behind the choice of (6.4) as the correct translation of (6.2), labelling it as 'obvious'.[6] One possibility, following [Halbach 2001a], is to take into account the modal status of sentences like (6.3). The deflationist seems to be happy to consider instances of (6.1) analytic;[7] analyticity should thus be preserved via translations. However, (6.2) is clearly analytic for a competent English speaker, whereas (6.3) is not analytic for all competent Italian speakers. An obvious way to perform a translation which preserves the modal status would thus be to go from (6.2) to (6.4).

The conclusion that truth cannot be a genuine predicate thus follows by considering the extensions of 'is true' and 'è vero': the class of objects to which 'is true' is ascribed must be different from the class of objects to which 'è vero' can be ascribed. But if 'è vero' attributes

[6] Cfr. [Strawson 1949, p. 84].

[7] We do not embrace a discussion of Lewy's argument, which doubts the necessity of the T-schema, and assume that the deflationist has successfully restricted its instances so to avoid Lewy's challenge.

a property to Italian sentences, its translation 'is true', to be legitimate, should ascribe the same property to the same class of objects. Thus 'è vero' and 'is true' cannot attribute any property to sentences.

To our purposes, the brief description of thesis (i) just given should suffice. We only hope that the reader finds the presence of (i) in our characterisation of deflationism now rooted in its historical development.

Thesis (ii), in its first component, reminds us that at the heart of the deflationary standpoint about truth lies the disquotational character of the truth predicate. Emphasising the disquotational role of the truth predicate, however, seems at odds with the views held by Strawson, Williams and the prosententialists: for them, in fact, instances of (6.1) are secondary uses of 'is true', and they judge anaphoric and *performative* speech acts such as 'That's true' in response of a declarative sentence uttered by another speaker as authentic uses. Moreover, the disquotational character of the truth predicate presupposes a quotation device which is usually available for sentences only.[8] According to the disquotationalist, in the end, there seem to be objects to which the truth predicate is applied.

The truth predicate, according to (ii), can be taken to be redundant in the case of instances of (6.1): to assert

(6.5) 'Lilies are flowers' is true

is equivalent to asserting

(6.6) Lilies are flowers

itself. How this equivalence has to be understood is beyond the scope of the present work: [Field 1994] explains it in terms of 'cognitive equivalence', a notion that closely resembles

[8] Although [Horwich 1998] proposes a quotation device defined on propositions.

analyticity. For a discussion of the modal status of the disquotation sentences and of Field's proposal we refer to [Halbach 2001c].

(ii) crucially prescribes the presence of non dispensable uses of 'is true'. If one is convinced by Xenophon and Plato's apologies holding that Socrates was neither corrupting the youth not committing impiety, and thus he had no reason to lie in his last trial in 399 BC, he may affirm:

(6.7) The opening sentence of Socrates' defence in his last trial is true.

Even if we ignore the sentence Socrates precisely uttered at the beginning of his speech, we perform a blind ascription using the locution 'is true'. Unlike (6.5), the truth predicate in (6.7) cannot be eliminated.

Moreover, as Quine noticed,[9], in order to assert an infinite number of sentences, we need the truth predicate to be able to secure a dimension of generality that cannot be captured by a single term. For instance, generalising on the single term 'Hypatia' in

(6.8) Hypatia is mortal

yields a sentence of the form

(6.9) All humans are mortal.

However, generalising on a singular term associated with the complex expression 'Hypatia thinks' in

(6.10) Hypatia thinks or it is not the case that Hypatia thinks

[9]See [Quine 1986, Quine 1990].

requires, according to Quine, the presence of a truth predicate. To assert in fact an 'infinite lot' of sentences of the form (6.10), where 'Hypatia thinks' has been replaced by a metavariable for names of sentences, in fact, calls for a generalisation on its syntactic form, that is we need to assert

(6.11) All sentences of the form 'φ or not φ' are true.

The truth predicate employed in (6.11) is again not eliminable: Quine refers to the possibility of accessing this dimension of generality with the emphatic expression 'semantical ascent'. In particular, he writes

> ...but if we want to affirm some infinite lot of sentences that we can demarcate *only by talking about the sentences* [A/N: my emphasis], then the truth predicate has its use. ([Quine 1986, p.12])

(ii) thus tells us that, besides blind ascriptions of the kind exemplified by (6.7), the truth predicate does not serve any purpose beyond the logico-linguistic one of making possible general claims that are not otherwise expressible.

Some authors have tried to make this Quinean motto precise: [Halbach 1999] has proposed to replace this 'infinite lots of sentences' with infinite conjunctions or disjunctions. Although other disquotationalists such as Field seem to explicitly talk about infinite conjunctions and disjunctions,[10] Quine seems to be only interested in uses of 'is true' in which we can capture a general claim about sentences *only by talking about sentences*, and infinite conjunctions and disjunctions do not satisfy this requirement. seems -> appears?

At any rate, the problem of making precise the sense in which the truth predicate is *exclusively* an indispensable tool to produce general claims is highly debated: as it has been pointed out,[11], for instance, it seems that any predicate allows generalisations that could not

[10] Cfr. [Field 1986, p. 57].

[11] Cfr. [Heck 2009, p. 47].

be possible without it. For instance, the sentence

(6.12) All axioms of ZFC contain the symbol '$\in$'

could not be formulated without the predicate 'contain the symbol $\in$'. Of course the deflationist would hold that 'is true' is *just* a device for generalisations, and this tenet cannot be dismissed so easily. What Heck's example seems to indicate, however, is that the deflationary that the truth predicate is only a linguistic tool to produce general claims cannot be fully explained by the possibility of expressing sentences like (6.11).

It is not our purpose to investigate the plausibility of the thesis (ii); we just wanted to introduce it and highlight the fact that, although it seems still to lack a precise rendering, it is an integral part of the deflationary conception. We now move to a brief description of thesis (iii).

Philosophers have argued that the notion of truth is required to explain the success of scientific theories: in other words, empirical prosperity of our best scientific theories can be explained in terms of their truth or approximate truth.[12] Deflationism, as characterised by theses (i)-(ii), is thus forced to justify the explanatory power of the notion of truth, if any, on the face of its insubstantiality.

[Horwich 1998, pp. 48-49] has argued that the deflationist, and the minimalist in particular, should not be worried about the explanatory role that the notion of truth seems to play in making sense of the prosperity of scientific theories. On the contrary, this role is compatible with (i) and (ii). Consider the theory

(6.13) Nothing goes faster than light.

In this case we have access to the theory simply by listing its axioms. Therefore the explana-

[12] Cfr. [Putnam 1978].

tion of the empiric success of the theory in terms of its truth

(6.14) The theory that nothing goes faster than light is successful because it is true,

is equivalent to

(6.15) The theory that nothing goes faster than light is successful because nothing goes faster than light.

that does not involve the notion of truth. Unsurprisingly, the truth predicate reveals itself to be an essential tool, according to Horwich, when from singular theories we move to general claims concerning theories whose content we might ignore. The sentence

(6.16) true theories yield accurate predictions,

which indeed makes essential use of the truth predicate to express a possibly infinite lot, is understood by Horwich as an instance of (ii), thus in line with the deflationist's understanding of the notion of truth.

It seems, however, that Horwich's brief remarks do not really explain how it is possible to retain the explanatory role of the truth predicate on the face of its insubstantiality. Or at least they do if one assumes a clear sense in which a deflationary notion of generalisation can be spelled out. But we have seen that there is no free lunch for the deflationist in that respect.

Another possibility, which seems also to be implicit also in Horwich's minimalism, is to explain the success of our beliefs in terms of other substantial properties. One obvious option is predictive power. A naturalistic explanation would thus affirm that successful scientific theories are the ones that increase our probability of knowing the outcome of a certain set of conditions or of certain manipulations of the physical matter.[13]

From the philosophical side, a crucial implication of (iii) is that epistemology has to be carried out without a substantial notion of truth. [Williams, M. 1986] for instance, defends this thesis. On the contrary, [Kirkham 1992] seems to doubt the possibility of giving up a substantial notion of truth, unless serious incisions in classical epistemology are made.

Again we do not intend to take a stance on the exegesis of (iii), although it seems a quite central and often overlooked theme. By showing that such a thesis is often matched with deflationism in the philosophical literature we fulfilled the goal of this introductory section.

Also, we are not worried from the apparent obscurity of our presentation of (i)-(iii). This obscurity must have been perceived also by other philosophers, when they tried to make more precise the actual commitments of deflationism, as partially depicted by (i)-(iii), by looking at the deductive power of deflationary acceptable systems for truth.[14]

6.2 Notational Intermezzo: Some Reflection Principles

Before getting into the details of the conservativeness argument against deflationism, it seems useful to recall some technical notions that will be employed in what follows and to mark the boundaries of the portion of mathemathematical reflection we are primarily interested in.

By accepting a theory O, we are committed to its soundness and consistency. For a suitable choice of O, Tarski's theorem tells us that the notion of truth for $\mathcal{L}_{\mathsf{O}}$-sentences is not definable in O itself. In order to express the soundness of O in $\mathcal{L}_{\mathsf{O}}$ we then usually resort to surrogates. If O contains a sufficient amount of arithmetic,[15] the uniform reflection

[13] Cfr. [Leeds 1978].

[14] As we did for the deflationist, we label 'inflationist' any fictional philosopher that agrees on the negation of some of (i)-(iii).

[15] We are assuming, from now on, that O contains elementary arithmetic EA, as presented for instance in [Beklemishev 2005], or $I\Delta_0(\mathsf{exp})$, as introduced in [Hájek&Pudlák 1993].

principle

$$\mathsf{RFN_O} \qquad \forall x(\mathsf{Bew_O}(\overline{\varphi(\dot{x})}) \to \varphi(x)),$$

—where $\overline{\varphi(\dot{x})}$ represents within O the result of formally substituting the variable x for the numeral for x in the formula φ, and $\mathsf{Bew_O}(x)$ a canonical provability predicate for O in $\mathcal{L}_\mathsf{O}$— represents the best approximation to a soundness claim for O available in $\mathcal{L}_\mathsf{O}$.

Over O suitably chosen,[16] $\mathsf{RFN_O}$ is stronger than its local version

$$\mathsf{Rfn_O} \qquad \mathsf{Bew_O}(\overline{\sigma}) \to \sigma$$

for all $\mathcal{L}_\mathsf{O}$-sentences σ, as O+($\mathsf{RFN_O}$) proves $\mathsf{Rfn_O}$ which in turn implies the canonical consistency statement $\mathsf{Con_O}$ defined in terms of $\mathsf{Bew_O}(x)$. Both soundness statements are thus properly stronger than O.

If conceptual resources for truth are available, one can capture in a single sentence the commitment to the soundness of O. The global reflection principle for O

$$\mathsf{GRP_O} \qquad \forall x(\mathsf{Sent}_{\mathcal{L}_\mathsf{O}}(x) \land \mathsf{Bew_O}(x) \to Tx)$$

appears to be the full soundness claim for O. In the presence of uniform disquotation sentences of the form (4.2), in fact, $\mathsf{GRP_O}$ derives $\mathsf{RFN_O}$.

As we have seen, the typed theory of truth CT[O] proves $\mathsf{GRP_O}$ for many choices of O. CT[O] is often considered an adequate theory of truth for O and thus it might be argued that the acceptance of CT[O] is implicit in the acceptance of O. However, the process does not have to stop here and we might wonder what is implicit in the acceptance of CT[O] itself. The iteration can be repeated through the transfinite: in the case in which O is PA, [Feferman 1991] has investigated the possibility of disclosing all commitments implicit in

[16] Cfr. previous footnote.

the acceptance of PA.

The outcome of this analysis is known as the *reflective closure* PA. If the step from PA to CT is admitted as natural, the subsequent construction of theories obtained from earlier ones via the addition of new (typed) truth predicates and respective compositional axioms can reasonably continue up to ε_0: this is in fact a natural halting point since Peano Arihtmetic can only prove that the representations of ordinal numbers within PA, the so-called ordinal notations, are well-ordered up to ε_0.

The resulting theory $\mathrm{RT}_{<\varepsilon_0}$[17] seems to disclose all commitments that are implicit in the choice of PA, if one considers justified only iterations of truth predicates and compositional axioms (but also of reflection principles) along orderings that can be proved to be well-orderings in the base theory, PA in this case.

[Feferman 1991] and [Fujimoto 2011] consider further iterations: it may also seem justified, once arrived at the stage $\mathrm{RT}_{<\varepsilon_0}$, to take into account iterations along well-orderings that are provable in $\mathrm{RT}_{<\varepsilon_0}$, and so on. The Feferman-Schütte ordinal Γ_0, the first strongly critical ordinal,[18] seems another natural halting point, as the theory $\mathrm{RT}_{<\Gamma_0}$ cannot prove that orderings equal or larger than Γ_0 are well-orderings. Although it is technically feasible also to consider further iterations, it seems that they would hardly be justifiable.

In what follows, more humbly, we will only be only be interested in the first step of the process of reflection just sketched. Philosophers have in fact focused their attention to the transition from PA to CT to criticise the alleged insubstantiality of the truth predicate advocated by the deflationist.

[17] See [Halbach 2011, p. 127] for a definition.
[18] Cfr. [Pohlers 2009, Ch. 2].

6.3 The Conservativeness Argument

According to the deflationary hallmark (iii), truth does not increase our knowledge of reality. [Shapiro 1998] mingled (i) and (iii) in a single, unexpectedly successful expression: according to the deflationist, truth is a *metaphysically thin* notion. Its fortune is quite surprising given its obscurity. Having said that, Shapiro should also be granted, together with [Horsten 1995] and [Ketland 1999], for a mathematical regimentation of its arcane characterisation: a deflationary acceptable (formal) theory of truth has to be conservative over its base theory.

Following [Halbach 2011], we call a sentence non-semantic if it does not involve the truth predicate or alternative semantic vocabulary.[19] Roughly, according to Horsten, Shapiro and Ketland, the deflationist can only accept a theory of truth—over a base theory O—which does not enable one to prove any new non-semantic consequences, that is no new theorems in the language of the base theory which were not provable in O already. Moreover, since adequate theories of truth—according to the standards set up by those authors, Ketland and Shapiro in particular—all prove new non-semantic consequences with respect to O, there is no such a thing as an adequate theory of truth which is also deflationary licit. In the next sections, we will examine more closely the adequacy criteria employed by Ketland and Shapiro.

In order to arrive at what we think is the best formulation of the conservativeness requirement for deflationary theories of truth, let us first recall some relevant axiomatisations of the truth predicate, already introduced in §2.2.2. We also refer to Chapter 2 for the notions of axiom schema and schematically axiomatised theory. They are formulated in the language $\mathcal{L}_{\mathsf{O}} \cup \{T\}$ where O is a theory interpreting a sufficient amount of arithmetic. The theories are displayed in Table 6.1. We denote with TB, UTB, $\mathsf{CT}\!\upharpoonright$ and CT the theories obtained in the way illustrated by Table 6.1 in the case in which O is PA.

[19] Precisely, not definable in the base theory: for instance, the arithmetical evaluation function or partial satisfaction predicates thus do not follow under this connotation of *semantic* if the base theory if PA.

O^T:=	O formulated in the language $\mathcal{L}_\mathsf{O} \cup \{T\}$ where schemata of O, if present, are extended to contain the truth predicate T.[a]
$\mathsf{TB}[\mathsf{O}]$:=	O^T together with the schema $T\ulcorner\sigma\urcorner \leftrightarrow \sigma$ for all $\mathcal{L}_\mathsf{O}$-sentences σ.
$\mathsf{UTB}[\mathsf{O}]$:=	O^T plus the schema $\forall x(T\ulcorner\varphi(\dot{x})\urcorner \leftrightarrow \varphi(x))$ for all $\mathcal{L}_\mathsf{O}$-formulas $\varphi(v)$.[b]
$\mathsf{CT}[\mathsf{O}]\upharpoonright$:=	O^T plus Tarski's inductive clauses turned into axioms but with schemata of O restricted to $\mathcal{L}_\mathsf{O}$.
$\mathsf{CT}[\mathsf{O}]$:=	O^T plus Tarski's inductive clauses turned into axioms.

Table 6.1: Some axiomatisations of typed of truth.

[a] O^T is obviously conservative over O.
[b] Here $\ulcorner\varphi(\dot{x})\urcorner := \mathsf{sub}(\ulcorner\varphi\urcorner, \ulcorner x\urcorner, \mathsf{num}(x))$, namely it denotes the result of formally substituting in $\varphi(x)$ the code of the variable x with the corresponding numeral.

The formulation of the argument. [Shapiro 1998] also starts with a theory A interpreting a sufficient amount of arithmetic. Then considers the addition of the truth predicate to its language and the theories $\mathsf{TB}[\mathsf{A}]$, $\mathsf{UTB}[\mathsf{A}]$, and $\mathsf{CT}[\mathsf{A}]\upharpoonright$. He first notices that all those axiomatisations of truth over A are conservative over it. A closer look at Shapiro's paper, however, clearly indicates that he has in mind the case in which A is PA.[20] Therefore, although his claim is true, Shapiro refers to a proof of the conservativity of $\mathsf{CT}[\mathsf{A}]$ by a simple model expansion that, for the case in which A is Peano arithmetic, is known to be flawed.[21] For a correct proof, see [Kotlarsky et alii 1981, Enayat&Visser 2013]. Then continues,

> Since the original language $\mathcal{L}$ contains ordinary quantification, one can state in $\mathcal{L}'$ that all the axioms of A are true and one can state in $\mathcal{L}'$ that the rules of inference preserve truth. Since these generalizations are obviously correct, an adequate theory of truth should have the resources to *establish* them. It follows,

[20] See for instance p. 499, second paragraph.
[21] This follows from Lachlan's theorem, mentioned in Chapter 4.

> or should follow, that *all the theorems are true* [...] So, if our deflationist takes A' to capture just about everything essential to truth concerning $\mathcal{L}$ and if the consequence relation is first-order, then our deflationist cannot say that all the theorems of A are true. ([Shapiro 1998, p. 498]

In the quotation, A' is ours $\mathsf{CT}[\mathsf{A}]\restriction$. It is useful at this stage to extract from the quotation the the following partial desideratum for adequate theories of truth:

Reflective Assumption. An adequate theory of truth should be able to state and establish the truth of the axioms of the base theory and the truth-preserving character of its rules of inference.

It is not clear from the context whether Shapiro considers TB and UTB as satisfying the Reflective Assumption, but he surely thinks that $\mathsf{CT}\restriction$ does.[22] In any case, as soon as the object theory A is schematically axiomatised, the theories $\mathsf{TB}[\mathsf{A}]$, $\mathsf{UTB}[\mathsf{A}]$ and $\mathsf{CT}[\mathsf{A}]\restriction$ do not satisfy the Reflective Assumption as they can only state and not prove that all axioms of A are true. In order to prove the claim 'all instances of induction of PA are true',[23] for example, we need an instance of the induction of PA extended to contain the truth predicate together with compositional axioms for truth, that is Tarski's inductive clauses. However, TB and UTB have the extended induction but do not have the compositional axioms, whereas $\mathsf{CT}\restriction$, which has the compositionality required by the argument, does not have the extended induction. Moreover, for $\mathsf{TB}[\mathsf{A}]$ and $\mathsf{UTB}[\mathsf{A}]$, by adapting the consistency proof given by Tarski for theories along the lines of TB,[24] we can prove that such a result is impossible: they cannot in fact prove any generalisation involving the truth predicate. At any rate these considerations, overlooked by Shapiro, seem to strengthen his requirements on the adequacy of the theories for truth he discusses. Already according to the partial adequacy desideratum represented by the *Reflective Assumption*, TB, UTB and $\mathsf{CT}\restriction$ are ruled out.

[22] See also [Shapiro 2002, p. 108] for a case in which Shapiro unequivocally considers PA as base theory.
[23] Cfr. Chapter 1.
[24] For a modern exposition, see [Halbach 2011, Theorem 7.7].

Ketland's formulation of the adequacy condition for theories of truth is more careful. According to him, the key detail which makes a theory of truth adequate in this context resides mainly in its capability of establishing the global reflection principle $\mathsf{GRP}_{\mathsf{O}}$ for the object theory O, what Ketland calls the 'equivalence principle'. [25]

We will see below that if the precedence is given to the possibility of reconstructing the metamamathematical reasoning behind the acceptance of O, then the theories of truth with 'disentangled' syntax introduce in Chapter 4 will provide us with a different strategy to reassemble some metatheoretic insights that are not available in O.

Incidentally, it is widely recognised that the provability of the Global Reflection Principle for the object theory is an important desideratum for a satisfactory theory of truth: truth theorists have recently tried to formulate lists of requirements to characterize the adequacy of a theory of truth,[27] and the provability of the 'truth' of the object theory always appears to be a desirable feature.

We should also add that the Global Reflection Principle is so appealing mainly because of the way in which it is proved: its proof in fact mimicks within the theory of truth the semantic consistency proof already considered by Tarski.[28] This is essentialy what Shapiro demanded for adequate theories of truth: reflecting reasoning leading to a soundness claim for O. We can then implement the *Reflective Assumption* in the following, more general, requirement:

[25] Some authors also seem to emphasise the importance of the provability of the Gödel sentence from the theory of truth. As Ketland himself puts it:

> We certainly "recognize" that a Gödel sentence G for T is true (on the assumption that T itself is true), but our knowledge of its truth does *not* obtain from correct formal derivations *within* the theory T to which it applies. ([Ketland 1999, p. 87])

Ketland notices that CT decides the Gödel sentence, and thus it enables one to formally capture some essential insights which are implicit in the acceptance of PA but not provable in it. The adequacy of CT as a theory of truth for PA thus resides in the possibility of 'significantly transcend' the theory TB—but also UTB and $\mathsf{CT}\!\upharpoonright$—by providing a deduction of the Gödel sentence.[26] The fact that theories like CT decide the Gödel sentence for PA should be counted as a consequence of their adequacy.

[27] Cfr. [Leitgeb 2007], [Feferman 2008] and [Halbach&Horsten 2014].

[28] See [Tarski 1956a, §3, Theorem 5].

Reflective Adequacy Condition. An adequate theory of truth for a theory O, once established that all axioms of O are true and that all the rules of inference specific to the formulation of O preserve truth, should also be able to establish that all theorems of O are true.

The last piece of the puzzle towards a general formulation of the conservativeness argument is probably the toughest to put into place. We need to justify the claim that the deflationist is committed to theories of truth that are conservative over their respective base theories. In a slogan:

Conservativeness Requirement. The deflationist can only endorse theories of truth that are conservative over the base theory.

[Shapiro 1998] argues that the transition from O to its nonconservative extension O', obtained by adding only axioms essential to truth to O^T, by restricting the logical possibility, also determines the addition of some semantic content. In particular, if a formula φ is not derivable in O but it is a theorem of $\mathsf{CT}[\mathsf{O}]$, then the set $\mathsf{O} \cup \{\neg\varphi\}$ is consistent, whereas $\mathsf{CT}[\mathsf{O}] \cup \{\neg\varphi\}$ cannot be consistent.

Axioms for truth are then responsible for the increase of the explanatory power of the theory in relation to the mathematical structure partially captured by the axioms of O, and thus the notion of truth cannot be insubstantial after all. One of the crucial assumptions upon which the argument relies is then that the explanatory role of the truth predicate, its 'metaphysical thinness', can be tied up with some notion of logical consequence, first-order logical consequence more specifically. Shapiro's main goal is to show that the deflationist who accepts the *Conservativeness Requirement* has to give up first-order logical consequence and to accept stronger notions. This, of course, if the conservativeness argument, once properly formulated, actually works.

Back to the *Conservativeness Requirement*, [Halbach 2001a] has argued that the deflationist should not commit herself to the conservativity of the theory of truth. One way to articulate this claim is by looking at the status of $\mathsf{GRP}_{\mathsf{PA}}$.

We have seen in §6.1 that the deflationist holds that the truth predicate is indispensable for 'expressing' general facts. If the theory of truth is formulated following the disquotational spirit, that is when we have T-sentences as axioms, [Halbach 1999] has shown that the sense in which the truth predicate allows for generalisations cannot coincide with the provability of the general claim at stake. 'Expressing' can instead be understood as 'replacing an infinite list of sentences with a single one', a sort of finite reaxiomatisation. For instance, PA+ all instances of $\mathsf{Rfn}_{\mathsf{PA}}$ proves the same arithmetical sentences as $\mathsf{TB} + \mathsf{GRP}_{\mathsf{PA}}$. Halbach's example seems to suggest that, given a background (deflationary acceptable) axiomatisation of the truth predicate, the single sentence σ capturing the infinite set S, when added to the base theory, should prove the same sentences in the base language as the base theory plus S.

Expressing certain generalisations in our theory of truth, in the deflationist's sense, does not only enable us to capture in a single sentence an infinite list; it enables us also to exhibit a certain commitment towards general facts. As Field puts it:

> Shapiro says that deflationist holds that truth is "metaphysically thin". I am not sure what this means, but one thing that it better not to mean is that we cannot use it to express important things inexpressible or not easily expressible without it, or that we cannot use it to make commitments about matters not involving truth beyond those commitments which we could make without it; for it is a clear part of deflationist doctrine that truth is *not* metaphysically thin in that sense. ([Field 1999, p. 534])

Field also seems to consider $\mathsf{CT}\!\restriction$ as deflationary licit. There is also no doubt that he would want to exhibit commitment in the conjunction of all theorems of PA. The following observation shows that Halbach's strategy is not applicable to Field's case:

Observation 2. $\mathsf{CT}\!\restriction + \mathsf{GRP}_{\mathsf{PA}} \not\equiv_{\mathcal{L}_{\mathsf{A}}} \mathsf{PA} + \mathsf{Rfn}_{\mathsf{PA}}$.

Proof. $\mathsf{CT}\!\restriction + \mathsf{GRP}_{\mathsf{PA}}$ is the same theory as $\mathsf{CT}[I\Delta_0]$, and it proves $\mathsf{Con}_{\mathsf{PA}+\mathsf{Rfn}_{\mathsf{PA}}}$. *qed.*

If 'committing oneself' to certain general facts not involving truth—such as the conjunction of all theorems of PA—by means of the truth predicate is equivalent to the assumption of the general claim as an axiom over an accepted axiomatisation of the truth predicate, then it is not clear how the deflationist, and Field in particular, can accept $\mathsf{CT}\restriction$ without losing interest in the issue of the conservativity of the theory of truth.

The discussion just sketched renders the following passage from [Field 1999] even more striking:

> ...there is no need to disagree with Shapiro when he says "conservativeness is essential to deflationism". [Field 1999, p. 536]

We have thus an example of a deflationist who tolerates the conservativeness requirement for deflationary acceptable theories of truth. This suffices for us to put aside further discussions on the engagement of the deflationist to the conservativity of the theory of truth, and reason *as if* the deflationist had accepted Shapiro's matching of the insubstantiality of the truth predicate with the conservativeness of the theory of truth over its base theory.

We finally state our formulation of the conservativeness argument.

Conservativeness Argument. The deflationist can only endorse theories of truth that are conservative over their base theories. However, an adequate theory of truth ought to be able to formalise the Tarskian semantic consistency proof for its object theory. Gödel phenomena thus force an acceptable theory of truth not to be conservative over its object theory. Therefore there are no adequate, deflationary licit theories of truth.

We are inclined to support the view that the role of the truth predicate is indeed indispensable for the articulation of some general claims; moreover, we sympathise with the idea that the deflationist cannot accept the conservativeness requirement and at the same time preserve the role of the truth predicate expressed by (ii).

We will see in what follows that theories of truth with 'disentangled' syntax represent an appealing framework for the deflationist: in that framework he can maintain that the truth predicate is essentially a device for expressing commitments not expressible without it, but also freely articulate some significant metatheoretic arguments of syntactic nature, such as the one leading to the establishment of the soundness of O.

6.4 Conservativeness and Syntax

Some formulations of the conservativeness argument do not completely settle the issue of identifying the kind of conservativity the inflationist is requiring. [Shapiro 1998] seems to offer a powerful formulation of the conservativeness argument which includes also the possibility for O to be the empty set. [Ketland 1999], on the other hand, focuses on the case in which O is PA, and seems to have in mind the conservativity of the theory of truth over the underlying mathematics. [Halbach 2001a] and [Shapiro 2002] on the other hand seem to focus on the conservativity of the truth theory over the underlying ontology of expressions. In the usual setting, therefore, their position is not different from Ketland's, in fact:

> ... what is a suitable theory of sentence types? There are good reasons for picking Peano arithmetic, although other theories might be suitable as well. Because the discussion so far has mainly focused on Peano arithmetic (which has some nice features as a base theory), I will concentrate on it; similar points, however, can be made for several other theories (like ZF). ([Halbach 2001a, p. 182])

In earlier chapters we touched on the difficulties related to the quest for a philosophically satisfactory and sufficiently tractable theory of sentence types. A crucial requirement on the underlying ontology of objects to which truth is attributed, like in the cases of PA and ZF, is to show that they possess an ontological structure similar to that of sentences, which are still the intended objects to which truth is ascribed. Peano arithmetic is often chosen as it is well-known that the structure of natural numbers is isomorphic to the structure of strings of a finite or countable alphabet.[29]

[29] Cfr. [Corcoran et alii 1974].

The contribution that we intend to offer to the discussion around the explanatory role of the truth predicate in the present chapter is not dependent on the details of the theory of truth bearers. We are principally concerned with a discussion of the status of theories of truth with disentangled syntax in this context regardless the specific choice of the theory of syntax. We will see that they will provide us with a framework in which it is possible to distinguish between conservativity of the theory of truth over the theory of the objects of truth and over the object theory.

Conservativeness, Syntax and Logic. As we remarked above, Shapiro's formulation of the conservativeness argument seems to embrace also the case in which the object theory O is pure logic. The conservativeness requirement thus implies that a theory of truth constructed over pure logic, call it $\mathsf{T[FOL]}$, ought to be conservative over FOL. [Halbach 2001a] ruled out this possibility by showing that already the addition of the T-sentences over FOL determines the provability in the resulting theory of the existence of two-objects, which is obviously not provable FOL. This showed that the conservativeness requirement—at least in Shapiro's first formulation—is at least overstated and that already very simple principles of truth carry with them some ontological weight.

We now show with the help of some simple examples that the strategy behind the construction of theories of truth with disentangled syntax, which will be recalled in some detail below, might look ontologically neutral even with respect to pure logic.[30] Consider in fact the theory $\mathsf{TBD[FOL^*]}$, formulated in a first-order language $\mathcal{L}^{**} = \mathcal{L}^* \cup C$ expanded with a countable set of constant symbols $C = \{c_1, c_2, ...\}$ that will act as names of $\mathcal{L}_{\mathsf{FOL}}$-sentences. $\mathcal{L}^*$ in turn has two-sorts of variables, one intended to range over the set the objects denoted by the constants in C, the other over the domain of some $\mathcal{L}_{\mathsf{FOL}}$-structure. A model $\mathcal{D}$ of $\mathsf{TBD[FOL]}$ would thus amount to a triple (M, N, I^+) where N can be taken to be a countable set of objects assigned to the elements of C by the interpretation I^+.

[30] Which by itself is not completely neutral, as unlike free logic it assumes the existence of one object.

It is easy to see that any model (M, I) of a set of $\mathcal{L}_{\mathsf{FOL}}$-sentences Γ can be expanded to a model of $\mathsf{TBD[FOL]}$. The key detail is that T-sentences now have the form $Tc_i \leftrightarrow \sigma_i$, where c_i is a constant symbol in C employed as a name, for instance, of the ith sentence in an enumeration of all sentences of $\mathcal{L}_{\mathsf{FOL}}$. The extension of the truth predicate is thus defined as the set of all objects in N which are the denotations of constants naming true $\mathcal{L}_{\mathsf{FOL}}$-sentences in (M, I). We notice that the stronger statement considered above can be strengthened to the case in which N contains only two objects.

Informally, the argument simply states that we can associate, by means of the T-sentences, each sentence of $\mathcal{L}_{\mathsf{FOL}}$ with a new constant in C and that the extension of the truth predicate is now defined on C, and not on the range of quantifiers of $\mathcal{L}_{\mathsf{FOL}}$. Therefore, the axioms of truth have trivially no impact on the domain of discourse of the object theory.

Already at this level, however, it might be useful to look at things from another perspective: let us denote variables ranging over N with k, l, m... and variables ranging over M with $x, y, z, \ldots$. The argument above shows that we cannot prove $\exists x \exists y\ x \neq y$ in $\mathsf{TBD[FOL]}$. On the other hand, we can certainly prove $\exists k \exists l\ k \neq l$ in $\mathsf{TBD[FOL]}$, that is the existence of at least two elements in N. So $\mathsf{TBD[FOL]}$ is conservative over FOL but already not conservative over the theory FOL^*, which is the two-sorted version of FOL just described.

The observations above offer a rather faithful picture of the structure of the philosophical scenario that can be extracted our 'disentanglement' of the syntax and that will surface again later: it in fact exhibits a Janus-faced structure in which the advantages for one side are always balanced with the possibility of a revenge argument from the other. For instance in the cases at hand the ontological neutrality of the T-sentences over pure logic has the drawback of displaying some impact on the ontology of expressions. In particular, the onerous version of the conservativeness argument given by Shapiro seems to be on the one hand rescuable by separating the domain of logic from the domain of truth bearers, but still if we focus our attention on the ontology of the objects of truth, the fulfillment of the conservativeness requirement over a simple theory such as FOL^*—which is admittedly a rather strange theory—is out of question.

In the sections to follow we abandon pure logic as base theory and consider object theories characterising some portions of mathematics.

Truth Theoretic vs. Mathematical. We have seen how some authors already chose conservativity over the theory characterising the underlying ontology of expressions as the right notion for the evaluation of the deductive power of the truth axioms. We have also mentioned the reasons that led them to consider Peano arithmetic as a suitable surrogate for a proper axiomatisation of strings of symbols.

The identification of natural numbers and sets with expressions is a standard procedure in metamathematical practice; it is suggested by Hilbert's program and it is the source of the limitative results which are the bases of contemporary mathematical logic. By advocating a formal setting in which syntactic and logical notions concerning the object theory are formalised in a different theory, however, we are not supporting a revision of logical work on truth theories. The framework presented is particularly suitable for the task of isolating patterns of metamethematical reasoning belonging to different categories that are relevant for our current philosophical concern. Theories of truth with disentangled syntax in this respect display the nice property of systematically disallowing interactions between domains of discourse belonging to different aspects of our metamathematical reasoning.

In the previous chapter we have provided a formal framework basen on these theories that was intended to to capture how syntactic objects on the one hand, and mathematical objects, although belonging to different areas of our metatheoretic reasoning, are usually related in practice. We now focus on the impact that this framework can have in the debate on the substantiality of the notion of truth

Let us now consider the familiar example of CT. It satisfies the reflective condition imposed by Shapiro, as $\mathsf{GRP}_{\mathsf{PA}}$ can be proved in it via a formalisation of the Tarskian semantic argument. Thus it is 'adequate' in the sense explained above. On the other hand, it is not an acceptable theory for the deflationist who tolerates the *Conservativeness Requirement*.

In his response to Shapiro,[31] Field tried to kill two birds with one stone, endorsing the conservativeness requirement but also denying that all the axioms employed in the proof of $\mathsf{GRP}_{\mathsf{PA}}$ in CT were 'essential to truth'. The extended induction axioms CT in fact, according to Field, rely on the well-foundedness of the structure of natural numbers and they are mathematical and not truth-theoretic in character. $\mathsf{CT}\restriction$ seems to be thus a better case to examine as it obtained by adding to PA only axioms characterising the notion of truth. Indeed, it is conservative over PA.[32]

We will shortly discuss the sense in which compositional axioms can be considered to be essentially truth-theoretic. We will also briefly recall why we think $\mathsf{CT}\restriction$ is not adequate according to the parameters discussed above. Let us first concentrate our attention on Field's understanding of the extended induction of CT.

To appreciate the emphasis put on the *mathematical* character assigned by Field to the extended induction axioms of CT, let us consider the following passage:

> There is nothing special about truth here [in the nonconservativity of CT (A/N)]: using any other notion not expressible in the original language we can get new instances of induction, and in many cases these lead to nonconservative extensions. ([Field 1999, p. 536])

As pointed out by [Heck 2009],[33] this quotation looks quite mysterious. On the one hand it would be natural to read 'not expressible' as 'not definable in the object theory'. In that case, however, already TB would suffice to define the set of true sentences of PA and the presence of the extended induction would not lead to nonconservative extensions. This drives Heck to conclude that compositional axioms are indeed essential to obtain the desired proof of $\mathsf{GRP}_{\mathsf{PA}}$ and that the extension of the schema of induction to the truth predicate cannot be seen as the only responsible for the nonconservativity of CT.

[31] Cfr. Field 1999.

[32] As with Shapiro's paper, the proof mentioned in [Field 1999] is not correct. An 'actual' proof of the conservativity of $\mathsf{CT}\restriction$ over PA is can be found in [Kotlarsky et alii 1981] and [Enayat&Visser 2013].

[33] A similar point is made in [Horsten 2011].

Also the claim that compositional axioms are 'essentially truth-theoretic' has to be clarified. There is no doubt, at least judging by the work of truth theorists, that any satisfactory account of the truth predicate should take into account how the truth predicate interacts with logical constants. They are thus essentially truth-theoretic in this straightforward sense.

If on the other hand we understand 'essentially truth theoretic' as meaning ontologically neutral, we notice that already the axiom governing the relation of the truth predicate with the negation symbol is nonconservative over FOL, as it again requires the objects of truth to be at least two. This indicates that compositional axioms cannot be intended as truth theoretic in this stronger sense, as they have some effect on the underlying ontology of truth bearers. In the usual setting in fact, where the contribution of syntactic principles is indistinguishable with the contribution of mathematical principles, compositional axioms can only be considered of 'mixed character', both mathematical and truth-theoretic.

It is only in a setting in which the syntax is separated from the mathematics that compositional axioms can be seen in their mixed truth-theoretic and syntactic character, which in turn does not bear any impact on the mathematical structure that models the object theory. An analog of an observation introduced above can be instructive: let $\underset{\cdot}{\neg}D[\mathsf{FOL}]$ be a theory in the two-sorted language $\mathcal{L}^{**} \cup \{T\}$ already presented above: $\mathcal{L}^{**}$ is the language expanding the language $\mathcal{L}_{\mathsf{FOL}}$ with a countable set C of new constants and featuring quantification over two separate domains, the domain of discourse of $\mathcal{L}_{\mathsf{FOL}}$ and the set of objects in C. We recall that names of $\mathcal{L}_{\mathsf{FOL}}$-sentences possibly belonging to the extension of the truth predicate live in a different domain with respect to the one over which standard variables of $\mathcal{L}^{*}$ range.

The axioms of $\underset{\cdot}{\neg}D[\mathsf{FOL}]$ are simply the axioms of FOL plus the corresponding axiom for the interaction of truth with negation, which says that if c is a name for $\neg\varphi$ and d for φ, then Tc if and only if $\neg Td$. Again this axiom over pure logic implies that the domain of the names of sentences of $\mathcal{L}_{\mathsf{FOL}}$ contains more than one object; on the other hand, it does not stipulate anything on the domain of quantifiers of the pure logical part of $\underset{\cdot}{\neg}D[\mathsf{FOL}]$.[34] This simple observation contributes to strengthen our belief that truth and syntax are deeply intertwined and even in the setting that will be proposed below, although it is possible to separate out the mathematical part of the theory of truth from the rest, it is much more difficult to isolate what is purely truth-theoretic from what is indeed syntactic.

A final remark concerning [Field 1999]. He seems to consider $\mathsf{CT}\restriction$ as adequate.[35] As we have noticed, however, $\mathsf{CT}\restriction$ does not have the resources to prove that all the axioms of PA are true. In other words, our 'adequate' theory of truth for PA does not even prove the truth of the pillars of our reasoning about natural numbers. This certainly does not seem to count in favour of $\mathsf{CT}\restriction$ and of Field's preferences.

Conservativeness and Truth Bearers. We now examine in some detail in what sense theories of truth with a 'disentangled' theory of syntax can be relevant for the discussion sketched above.

Let us take into account the theory $\mathsf{CTD}[\mathsf{O}]$, as presented in Chapter 4. We briefly recall its main components. $\mathsf{CTD}[\mathsf{O}]$ is formulated in the first-order, three-sorted language $\mathcal{L}_{\mathsf{TD}}$. Its quantifiers range respectively over elements of the domain $|\mathcal{M}|$ of a model $\mathcal{M}$ of O, which stands for the mathematical object theory; over the domain a model $\mathfrak{H}$ the theory HF of hereditarily finite sets,[36] and over a set of sequences in $|\mathcal{M}|$, living in a separate domain.

[34] Calling the additional axiom of $\underset{\cdot}{\neg}D[\mathsf{FOL}]$ (over logic) an 'axiom of negation' may seem to be excessive, as nothing really tells us that the function symbol $\underset{\cdot}{\neg}$ can be thought as the function expressing the concatenation of the negation symbol with the name of a formula. At any rate, the simple argument is as follows:

$\forall x, y\; x = y$	Absurdum Hypothesis
$c = \underset{\cdot}{\neg}c$	Hypotesis
$T\underset{\cdot}{\neg}c \leftrightarrow Tc$	Logic
$Tc \leftrightarrow \neg Tc$	$\underset{\cdot}{\neg}D[\mathsf{FOL}]$

The last two lines give the desired contradiction.

[35] Cfr. [Field 1999, p. 535].

[36] Cfr. §3.2.

We still use respectively $x, y, z, ..., k, l, m, ...$ and $a, b, c, ...$ to denote the corresponding variables.

The axioms of $\mathsf{CTD[O]}$, besides the axioms of O and the basic axioms of HF, comprehend compositional axioms for a binary satisfaction predicate Sat which applies to set theoretic codes of formulas of the language of O represented in the theory HF and to a sequence of variable assignments a; an axiom stating that the domain of sequences is nonempty and that given a sequence a we can always transform it into a sequence b in which the value of one element only has been changed;[37] finally, and most importantly, the axiom scheme of induction of the theory HF still applies to variables ranging over the domain of hereditarily finite sets but it is open to formulas containing quantification over all three domains and possibly parameters from all sorts of variables. We called this extended axiom schema $\mathsf{HF3}^*$. (Cfr. Table 4.1).

Crucially, $\mathsf{HF3}^*$ enables us to perform inductive reasoning involving the notion of truth *about* O: for instance, we can reason by induction on the logical complexity of the formulas of $\mathcal{L}_\mathsf{O}$ or by induction on the length of the derivations in O allowing into the induction formulas containing semantic vocabulary. We recall that $\mathsf{GRP}^s_\mathsf{O}$ and $\mathsf{Con}^s_\mathsf{O}$ denote respectively the (fixed, canonical) Global Reflection Principle for O and consistency statement for O formulated by means of the 'disentangled' representation of syntactic notions concerning O within HF as carried out in detail in §3.3, whereas their (fixed, canonical) counterparts in the language of O will be denoted by $\mathsf{GRP}^o_\mathsf{O}$ and $\mathsf{Con}^o_\mathsf{O}$. Let us also recall some relevant theories and definitions, presented already in Chapter 4 and 5. They are listed in Table 6.2.

Table 6.3 summarises the relationships between theories of truth with disentangled syntax just introduced and the syntactic and mathematical consistency and global reflection assertions. The theories $\mathsf{CTD[O]}$ and $\mathsf{CTD[O]}+\mathsf{AxT}^s_\mathsf{O}$ fulfill what we called the *Reflective Assumption*, although in the case of $\mathsf{CTD[O]}+\mathsf{AxT}^s_\mathsf{O}$, in a very peculiar way: the truth of all axioms of the schematic base theory O is in fact postulated on the basis of the informal

[37] This axiom is essentially due to [Craig & Vaught 1958].

$\mathsf{AxT}^s_{\mathsf{O}}$:=	$\forall k(\mathsf{Ax}^s_{\mathsf{O}}(k) \rightarrow \forall a\, \mathsf{Sat}(a,k))$;
H·O:=	the two-sorted theory in the language $\mathcal{L}_{\mathsf{H\cdot O}}$ featuring quantification on the syntactic and on the mathematical domain and whose axioms are the axioms of O and basic axioms of HF together with the schema of induction of HF extended to arbitrary formulas of $\mathcal{L}_{\mathsf{H\cdot O}}$.
$\mathsf{CTD[O]}^+$:=	with O schematically axiomatised, the theory CTD[O] with the schemata of O expanded to the entire vocabulary of $\mathcal{L}_{\mathsf{TD}}$;
$\mathsf{CTD[O]}^B$:=	the theory CTD[O] expanded with coding axioms forcing a bijection between 'mathematical' and 'syntactic' objects.

Table 6.2: Some theories of truth with disentangled syntax.

assumption of the soundness of O and not derived in the theory of truth. For the moment being, however, we will just take the adequacy in the sense of the *Reflective Assumption* of the theory $\mathsf{CTD[O]}+\mathsf{AxT}^s_{\mathsf{O}}$ as given and postpone a discussion of the status of the postulate $\mathsf{AxT}^s_{\mathsf{O}}$.

CTD[O] and $\mathsf{CTD[O]}+\mathsf{AxT}^s_{\mathsf{O}}$ can establish the truth of all theorems of O, as HF3* enables us to perform the inductive step that from the truth of the axioms of O and from the derivation of the truth preserving character of its rules of inference leads to the truth of all theorems of O. In this sense, they can capture our reflective reasoning leading to the explicit assertion of the soundness of O, which is not expressible in O itself. In particular, $\mathsf{GRP}^s_{\mathsf{PA}}$ seems thus to represents a general claim compatible with the deflationary hallmark (ii) no less than its mathematical counterpart $\mathsf{GRP}^o_{\mathsf{PA}}$. In the usual setting, however, once $\mathsf{GRP}^o_{\mathsf{PA}}$ becomes provable in the theory of truth, we lose another seemingly important requirement, its conservativity over the base theory, over O in this case. Therefore, in the usual setting,

O is finitely axiomatised :	$\mathsf{CTD[O]} \vdash \mathsf{GRP}^s_O$;	$\mathsf{CTD[O]} \vdash \mathsf{Con}^s_O$;
	...[a]	$\mathsf{CTD[O]}^{ca} \vdash \mathsf{Con}^o_O$;
O is schematically axiomatised:[b]	$\mathsf{CTD[O]} \nvdash \mathsf{GRP}^s_O$;	$\mathsf{CTD[O]} \nvdash \mathsf{Con}^s_O$;
	$\mathsf{CTD[O]} + \mathsf{AxT}^s_{\mathsf{O}} \vdash \mathsf{GRP}^s_O$;	$\mathsf{CTD[O]} + \mathsf{AxT}^s_{\mathsf{O}} \vdash \mathsf{Con}^s_O$;
	...[c]	$\mathsf{CTD[O]}^{ca} + \mathsf{AxT}^s_{\mathsf{O}} \vdash \mathsf{Con}^o_O$.

Table 6.3: Some results on theories of truth with disentangled syntax.

[a] The three dots witness that the addition of coding axioms to CTD[O] does not suffice to prove the mathematical version of the Global Reflection Principle for O.
[b] The statement is not quite precise: the result was proved only for the cases in which O is PA or ZF, but the same strategy is likely to work in other similar cases.
[c] Cfr. footnote *a*.

by fulfilling (ii) we are forced to give up (i) and (iii): the truth predicate seems to display some substantial explanatory power.

Surprisingly enough, theories of truth with 'disentangled' theory of syntax seem to open a way out for the deflationist. The results displayed in Table 6.4 indicate in fact that the notion of truth as characterised by those theories appears to be deflationary satisfactory also in the sense of theses (i) and (iii). The theories CTD[O] and CTD[O]+$\mathsf{AxT}^s_{\mathsf{O}}$ are in fact conservative over the mathematical theory O. In particular, any model of O can be expanded to a model of CTD[O] and CTD[O]+$\mathsf{AxT}^s_{\mathsf{O}}$. In the specific case considered above, CTD[PA]+$\mathsf{AxT}^s_{\mathsf{O}}$ can formalise the reasoning behind our commitment to the soundness of PA without exceeding PA itself in strength.

The result seems to support Field's thesis that the non conservativity of the theory CT relies upon a *mathematical use* of the induction schema of CT, although in a quite peculiar sense. In particular, the provablity the truth of all axioms of PA essentially relies on the possibility of proving the truth of all instances of induction of PA by employing one single

	O is finitely axiomatised	O is schematically axiomatised
CTD[O]:	✓	✓
CTD[O]+$\mathsf{AxT}^s_{\mathsf{O}}$:	✓	✓
CTD[O]$^+$:		×
CTD[O]^B:	×	
CTD[O]^B+$\mathsf{AxT}^s_{\mathsf{O}}$:		×

Table 6.4: Conservativity over the mathematical object theory O.

instance of the extended induction of CT. In order to establish that all instances of the induction of PA are true, in fact, we need to formally establish that, given an arbitrary formula $\varphi(v)$ of $\mathcal{L}_{\mathsf{O}}$ with only v free, if it is true when v is replaced by 0, and for all numerals $\overline{n}$, if the truth of $\varphi(v)$ when $\overline{n}$ replaces v entails the truth of $\varphi(v)$ when $S\overline{n}$ replaces v, then $\varphi(v)$ is true at each substitutional instance.

A quick look at the construction of CTD[PA]+$\mathsf{AxT}^s_{\mathsf{O}}$ already shows that this use of the induction schema of PA is not allowed. HF3* in fact, which is the induction schema of our theory of truth, cannot apply to 'mathematical' variables—or of sort o, borrowing some terminology from earlier chapters. More formally, the induction needed in CT to establish the truth of all instance of the induction applies in fact, as we have pointed out §2.3.1, to formulas of the language $\mathcal{L}_{\mathsf{A}} \cup \{T\}$ of the form $T\mathsf{sub}^o(\overline{\varphi}, \overline{v}, \mathsf{num}^o(x))$, where $\mathsf{sub}^o(\overline{\varphi}, \overline{v}, \mathsf{num}^o(x))$ represents in PA a primitive recursive function that applied to a code $\overline{\varphi}$ of a formula with one free variable returns the code of the result of formally substituting the code $\overline{v}$ of its free variable with the code of the numeral associated with the number x. Even if we have at our disposal a similar substitution function in HF,[38] it simply cannot tell

[38] Cfr. §3.2

us anything about numerals in the sense of the object theory PA. Thus $\mathsf{CTD[PA]{+}AxT}^s_{\mathsf{O}}$ cannot prove that all instances of the induction schema of PA are true.

Arguments like the one just sketched are different in nature from meta-theoretic arguments *about* PA intended as a specific formal system, such as the induction on the length of the derivations in PA that from the *Reflective Assumption* enables one to conclude that all the theorems of PA are true. In $\mathsf{CTD[PA]{+}AxT}^s_{\mathsf{O}}$ we can in fact show, for all ordinals n,

(H) : for all sequences a, the existence of a sequence (in the sense of HF) s of length $m < n$ such that $\mathsf{B}^s_{\mathsf{PA}}(s, m, k)$ and $\mathsf{Sat}(a, k)$ for an arbitrary formula k, together with the assumption of the existence of a sequence s' of length n such that $\mathsf{B}^s_{\mathsf{PA}}(s', n, l)$ for l again arbitrary, implies that $\mathsf{Sat}(a, l)$.

(C) : for all sequences a, for all k coding some $\mathcal{L}_A$-formula, the existence of a s of length n such that $\mathsf{B}^s_{\mathsf{PA}}(s, n, k)$ implies that $\mathsf{Sat}(a, k)$.

$\mathsf{HF3}^*$ suffices to capture this sort of reasoning, therefore $\mathsf{CTD[PA]{+}AxT}^s_{\mathsf{O}}$ is suitable to formalise the semantic consistency proof for PA and thus it is adequate in the sense of our *Reflective Adequacy Condition*.

However, if only arguments compatible with $\mathsf{HF3}^*$ are within the reach of our theory of truth, as the construction of $\mathsf{CTD[PA]{+}AxT}^s_{\mathsf{O}}$ shows, it will still remain conservative over the underlying mathematical theory PA. The conservativity will only be lost if some sort of interaction between mathematical and syntactic versions of the axiom schemata of the object theory and the syntax theory are allowed. The case of $\mathsf{CTD[PA]}^+$ is illuminating: as soon as the induction of PA is extended to the entire vocabulary of $\mathcal{L}_{\mathsf{CTD}}$, that is to formulas containing quantification and parameters from the other sorts and the satisfaction predicate, we can prove the theory of truth to be nonconservative over the mathematics, as it can prove a (fixed, canonical) consistency statement $\mathsf{Con}^o_{\mathsf{PA}}$ for PA formalised in PA itself.[39]

[39] Cfr. Theorem 4.4.

As we have seen in the previous chapter, the same result is obtained if we enlarge the theory of truth with axioms that determine a bijection between the domain of hereditarily finite sets and natural numbers. In the case of PA, this has been done by introducing a binary predicate applying to a 'syntactic' variable and a variable of the mathematical PA and axiomatising it in the spirit of the Ackermann interpretation of finite sets into natural numbers.

One of the points that we want to extract from these somewhat technical considerations is that Field's reply to Shapiro—in which he denied the exclusively truth-theoretic nature of the axiom schema of induction and emphasised its mathematical nature—as it stands, is rather not intelligible or, more bravely, simply wrong. On the one hand in fact we have seen that no axiom for truth, already in the case of the T-sentences, can be considered *purely* truth-theoretic, as it exhibits some impact on the underlying structure of truth bearers: therefore if understood in this sense Field's emphasis on the mathematical character of the induction of CT seems trivial. On the other, as Heck's remark made clear, stressing too much the role of the extended induction of CT in the semantic consistency proof of PA also appears unjustified.

By looking at CTD[PA], however, one can recover a point of view in which 'essentially truth theoretic' can be read as 'truth theoretic *and* syntactic' without suggesting any relation with the mathematical structure proper of the object theory. We can then say that the induction schema $\mathsf{HF3}^*$ of the syntactic part of CTD[PA], which enables us to formulate syntactic arguments *about* PA, crucially plays a *syntactic* role in our metamathematical reflection. The induction schema HF3* of CTD[PA] is thus not merely dependent on the nature of truth. Only in this distinguished sense, we affirm that Field's reply to Shapiro's challenge captures some important aspects of the reflective reasoning leading to the nonconservativity of adequate theories of truth.

Furthermore, theories of truth with disentangled syntax, by filtering out the components (of the first step) of our metamathematical reflection, offer to the inflationist as well the possibility of rephrasing her position from an original point of view. We remark once more how already in the early stages of the debate the conservativity of the theory of truth over the underlying theory of truth bearers has been taken into account; in fact by the inflationist's side:

> So, at best, the proper conclusion of my above-quoted little argument is that a deflationist should hold that truth is conservative over the underlying syntax. ([Shapiro 2002, p. 110])

Translated into the new vocabulary offered by theories of truth with disentangled syntax the inflationist would probably suggest to weigh up the impact that the theory of truth has over our new ontology of expressions which is now detached from the domain of discourse of the object theory. In other words, the inflationist is likely to suggest to consider $\mathsf{H}{\cdot}\mathsf{O}$ as our new base theory and rephrase our *Conservativeness Argument* accordingly. From this perspective the situation immediately appears less comfortable for the deflationist. The relevant results are summarised in Table 6.5.

	O is finitely axiomatised	O is schematically axiomatised
$\mathsf{CTD}[\mathsf{O}]$:	$\times$	$\checkmark$[a]
$\mathsf{CTD}[\mathsf{O}]{+}\mathsf{AxT}^s_{\mathsf{O}}$:	$\times$	$\times$
$\mathsf{CTD}[\mathsf{O}]^+$:		$\times$
$\mathsf{CTD}[\mathsf{O}]^B$:	$\times$	
$\mathsf{CTD}[\mathsf{O}]^B{+}\mathsf{AxT}^s_{\mathsf{O}}$:		$\times$

Table 6.5: Conservativity over the two-sorted theory $\mathsf{H}{\cdot}\mathsf{O}$.

[a] We have only a plausibility argument for this result, and not a full proof, which relies heavily on the choice of O.

If O is finitely axiomatisable, then $\mathsf{CTD}[\mathsf{O}]$ is not conservative over $\mathsf{H}{\cdot}\mathsf{O}$, as it proves $\mathsf{Con}^s_{\mathsf{O}}$. Similarly, the theory $\mathsf{CTD}[\mathsf{O}]{+}\mathsf{AxT}^s_{\mathsf{O}}$ proves $\mathsf{Con}^s_{\mathsf{O}}$ when O is schematically axiomatised. The theories $\mathsf{CTD}[\mathsf{O}]^+$, $\mathsf{CTD}[\mathsf{O}]^B$ and $\mathsf{CTD}[\mathsf{O}]^B{+}\mathsf{AxT}^s_{\mathsf{O}}$, on the contrary, for their specific construction, behave in a uniform way and prove a suitable consistency statement in the language of O for O which is not provable in O itself. This seems to support the thesis that truth has indeed an explanatory role concerning non semantical facts about the underlying ontology of expressions, even if we separate out the objects of truth from the objects representing the domain of discourse of our underlying mathematics.

A closer inspection of Table 6.5, however, also reveals some possible worries for this inflationist interpretation of the results presented: if O is schematically axiomatisable, we can only prove $\mathsf{CTD}[\mathsf{O}]$ to be nonconservative over $\mathsf{H}{\cdot}\mathsf{O}$ if we postulate the syntactic—i.e. in the language $\mathcal{L}_{\mathsf{HF}}$—assertion of the truth of all axioms of O $\mathsf{AxT}^s_{\mathsf{O}}$. This is reinforced by Theorems 4.1 and 4.2 which show that $\mathsf{Con}^s_{\mathsf{ZF}}$ and $\mathsf{Con}^s_{\mathsf{PA}}$ are not provable in $\mathsf{CTD}[\mathsf{ZF}]$ and $\mathsf{CTD}[\mathsf{PA}]$ respectively.

It might be thought that the assumption of $\mathsf{AxT}^s_{\mathsf{O}}$ is innocuous, as the uniform Tarski biconditionals for O are provable in $\mathsf{CTD}[\mathsf{O}]$, and thus every time that we face an axiom of O, or an instance of the axiom schemata of O, we can always prove it true by reflecting on it via those biconditionals. Moreover, the addition of an axiom such as $\mathsf{AxT}^s_{\mathsf{O}}$ is characteristic of the strategy behind all conservativity proofs contained in Table 6.4, based on the technique of constructing a finitely axiomatised extension of any recursively axiomatisable theory by means of a satisfaction predicate and a weak theory of assignments.[40]

As we have seen, however, proving the general claim stating the truth of all the axioms of O, especially when O is schematically axiomatised, seems to be one of the main requirements for a satisfactory theory of truth, at least according to what we called *Reflective Assumption*. This requirement seems in fact to precede the discussion of the cogency of the conservativeness requirement for deflationary acceptable theories; and the fact that we can construct a

[40] Cfr. [Craig & Vaught 1958].

finitely axiomatisable extension by resorting to $\mathsf{AxT}^{s}_{\mathsf{O}}$ and a satisfaction predicate does not necessarily imply that the resulting theory is adequate in the sense explained above.

A solution for the inflationist is to stress that if a theory does not satisfy the *Reflective Assumption*, then it is simply not adequate. Moreover, also by deflationary means, postulating a general claim such as the truth of all axioms of O, when O is schematically axiomatised, surely gives us the grounds to commit ourselves to the infinite conjunction of all the axioms of O, and this commitment is something that either the inflationist or the deflationist surely want to express. Since all the theories which mark an × in Table 6.5 with respect to conservativity over $\mathsf{H{\cdot}O}$ are adequate in this specific sense, one might be safe in concluding that truth *does indeed* have an explanatory role with respect to the underlying ontology of expressions, even though in a finer-grained sense than what displayed by the usual setting.

Before extracting some concluding remarks from the considerations above, we address a natural objection that the arguments just sketched could engender. Does the transition from O to $\mathsf{H{\cdot}O}$ trivialise the project of the disentanglement of the syntax? In other words, if the ontology of expressions is again part of the structure characterised by the object theory, aren't we pulled back in the old debate?

In response of this objection, it would be sufficient to highlight few insights offered by theories of truth with 'disentangled' syntax. The objection seems to stress the fact that, on the one hand, theories in the style of $\mathsf{CTD[O]}$ are too artificial to draw any significant conclusion; on the other, it highlights that, when one tries to render $\mathsf{CTD[O]}$ less artificial, then everything collapses into the good old setting. By remarking *some*, even minimal, achievements of our theories in relation to the debate of the substantiality of truth, in fact, it seems that we would at least save our setting from the threat of those arguments 'from trivialisation'.

One achievement is surely a response to Field's understanding of the schemata of the theory of truth. As far that the author knows, in fact, there has been no response to Field's argument, before Heck's unpublished paper, which emphasised the difference between schemata of the theory of truth and schemata of the object theory.

Another achievement is the formulation of a further notion of conservativity over the theory of the objects of truth, which accompanies the already known notions of conservativity over the mathematical theory and over pure logic.

A third, related, achievement is the accent put on the fact that if one wants to evaluate the impact that the truth predicate has on the explanation of some (meta)mathematical facts, there are really three components to be isolated: there is the truth predicate (or equivalent semantic machinery), there is a theory of the objects of truth, usually sentence-types, and there is a mathematical theory which represents the subject matter, the universe over which these sentences are interpreted. Our setting makes clear one thing: if semantic machinery can sometimes be detached from the mathematical object theory, there is no way to separate it from the syntactic theory. Truth and syntax seems to be deeply intertwined, and it is extremely hard to consider axioms for truth, even of the most basic form, *purely* truth-theoretic.

Even from a pessimistic point view, it seems, the scrutiny of theories of truth with 'disentangled' syntax can, after all, provide some useful insights.

6.5 Conclusions

All axioms characterising the truth predicate seem to have a twofold truth-theoretic and syntactic nature. Whereas it is possible to isolate the mathematical content of our theory of truth from the other components, clearly distinguishing between the truth theoretic content from the syntactic content is not a viable option.

Theories of truth with disentangled syntax enable us to discriminate between two types of conservativity of the theory of truth: conservativity over the theory of the objects of truth and conservativity over the mathematical object theory. Unless bridge principles between the syntactic and the mathematical domain are added, those theories are conservative over the mathematical object theory. On the contrary, all theories that satisfy what we labeled *Reflective Assumption* are not conservative over the theory of truth bearers.

Even if in a refined sense, therefore, the theory of truth seems to license new syntactic consequences, enhancing our possibility of formulating metamathematical arguments about the object theory O.

There is still one unresolved issue. Generalisations in the style of the syntactic claims GRP^s_O seem to be perfectly in line with the deflationary understanding of the truth predicate. They are in fact syntactic claims expressing our implicit commitment to the soundness of the theory O. If the deflationist accepts the conservativeness requirement, he cannot even accept GRP^s_O as a consequence of a deflationary licit theory of truth, or even postulate it as an axiom. The argument seems still to push the deflationist to give a reasonable explanation of what a commitment towards a general claim could be.

Bibliography

[Ackermann 1937] Ackermann, W. (1937), 'Die Wiederspruchsfreheit der allgemeinen Mengenlehre', *Matematische Annalen* 114: 305–315.

[Austin 1950] Austin, J. L. (1950), 'Truth', *Proceedings of the Aristotelian Society*, 24: 111-128.

[Barwise 1975] Barwise, J. (1975), *Admissible Sets and Structures*, Springer, Berlin.

[Beklemishev 2005] Beklemishev, Lev D. (2005), 'Reflection principles and provability algebras in formal arithmetic', *Russian Mathematical Surveys*, 60, 197–268.

[Benacerraf 1973] Benacerraf, P. (1973), "Mathematical truth", *Journal of Philosophy* 70: 661-679.

[Betti 2008] Betti, A. (2008), 'Polish Axiomatics and its Truth: On Tarski's Leśniewskian Background and the Ajdukievicz Connection', in [Patterson 2008, pp 44-71].

[Boolos 1993] Boolos G. (1993), *The Logic of Provability*, Cambridge University Press.

[Boolos 1999] Boolos, G. (1999), *Logic, Logic and Logic*, Harvard University Press.

[Boolos et alii 2007[5]] Boolos, G., Burgess J.P. and Jeffrey R. (2007), *Computability and Logic*, Cambridge University Press.

[Burgess&Rosen 1997] Burgess, J. and G. Rosen (1997), *A subject with no object. Strategies for a nominalistic interpretation of mathematics*, Clarendon Press, Oxford.

[Burgess&Burgess 2011] Burgess, A. and J. P. Burgess (2011), *Truth*, Princeton University Press.

[Carnap 1934] Carnap, R. (1934), *Logische Syntax der Sprache*, translated by A. Smeaton as *The Logical Syntax of Language*, London: Kegan Paul, Trench, Trubner & Co.

[Chang&Keisler 1990] Chang, C. C., J. Keisler (1990[3]), *Model Theory*, North-Holland.

[Hodges&Chiswell 2007] Chiswell, I. and W. Hodges (2007), *Mathematical Logic*, Oxford University Press.

[Corcoran et alii 1974] Corcoran J., F. William and M. Maloney (1974), "String Theory", *Journal of Symbolic Logic*, 39,4.

[Corcoran 2012] Corcoran, J. (2012), 'Schema', *The Stanford Encyclopedia of Philosophy* (Summer 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2012/entries/schema/>.

[Craig & Vaught 1958] Craig, W. and W. Vaught (1958), 'Finite Axiomatisability Using Additional Predicates', *The Journal of Symbolic Logic* 23: 289-308.

[Crivelli 2004] Crivelli, P. (2004), *Aristotle on Truth*, Cambridge University Press.

[Cutland 1980] Cutland, N. J. (1980), *Computability. An introduction to recursive function theory*, Cambridge University Press.

[Davidson 1967] Davidson, D. (1967), 'Truth and Meaning', *Synthese*, 17: 304Ð23; reprinted in [Davidson 2001].

[Davidson 1969] Davidson, D. (1969), 'Truth to the facts', *Journal of Philosophy*, 66: 748-764.

[Davidson 2001] Davidson, D. (2001), *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, (2nd edn).

[Davis 1965] Davis, M. (ed.), *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems, Computable Functions*, Hewlett, New York, Raven Press.

[Devlin 1984] Devlin, K. (1984), *Constructibility*, Springer.

[Enayat&Visser 2013] Enayat, A. and A. Visser (2013), *New Constructions of Satisfaction Classes*, Logic Preprints Series, University of Utrecht.

[Enderton 2001] Enderton, H. (2001), *A Mathematical Introduction to Logic*, 2nd ed. (New York: Harcourt/ Academic Press).

[Feferman 1960] Feferman, S. (1960), 'Arithmetization of metamathematics in a general setting', *Fundamenta Mathematicae*, 49, 35-91.

[Feferman 1962] Feferman, S. (1962), 'Transfinite recursive progressions of axiomatic theories', *Journal of Symbolic Logic* 27: 259-316.

[Feferman 1982] Feferman, S. (1982), "Inductively presented systems and the formalization of metamathematics", in Van Dalen et alii (ed.) (1982), *Logic Colloquium '80*, North-Holland, Amsterdam.

[Feferman 1989] Feferman, S (1989), “Finitary inductively presented logics”, in Ferro, R. et alii (eds.) (1989), *Logic Colloquium 1988*, North-Holland, Amsterdam.

[Feferman 1991] Feferman, S. (1991), “Reflecting on Incompleteness”, *Journal of Symbolic Logic*, 56, 1-49.

[Feferman 1998] Feferman, S. (1998), *In the Light of Logic*, Oxford University Press, Oxford-New York.

[Feferman 2008] Feferman, S. (2008), ‘Axioms for determinateness and truth’, *The Review of Symbolic Logic* 1, 2: 204-217.

[Ferreira 2005] Ferreira, F. (2005), ‘A Simple Proof of Parsons’ Theorem’, *Notre Dame Journal of Formal Logic*, 46(1).

[Field 1972] Field, H. (1972), ‘Tarski’s theory of truth’, *Journal of Philosophy* 69: 347-375.

[Field 1980] Field, H. (1980), *Science without numbers*, Princeton University Press.

[Field 1986] Field, H. (1986),‘The deflationary conception of truth’, in MacDonald G. and C. Wright (eds.)*Fact, Science and Morality: Essays in A.J. Ayer’s* Language, Truth and Logic, Blackwell, Oxford and New York.

[Field 1994] Field, H. (1994), ‘Disquotational truth and factually defective discourse’, *The Philosophical Review*, 103: 405–452.

[Field 1999] Field, H. (1999), “Deflating the conservativeness argument”, *Journal of Philosophy* 96 (10):533-540 (1999).

[Fitting 2007] Fitting, M. (2007), *Incompleteness in the Land of Sets*, College Publications, London.

[Franzel 2004] , Franzel, T. (2004), *Inexhaustibility. A non-exhaustive treatment*, Lecture Notes in Logic 16, ASL.

[Frege 1892] Frege, G. (1892), ‘Über Sinn und Bedeutung’, in Zeitschrift f§r Philosophie und philosophische Kritik, 100: 25Ð50. Translated as ÔOn Sense and ReferenceÕ by M. Black in *Translations from the Philosophical Writings of Gottlob Frege*, P. Geach and M. Black (eds. and trans.), Oxford: Blackwell, third edition, 1980.

[Fujimoto 2011] Fujimoto, K. (2011), ‘Autonomous progression and transfinite iteration of self-applicable truth’, *Journal of Symbolic Logic* 76 (3): 914–945.

[Fujimoto 2012] Fujimoto, K. (2012), ‘Classes and Truths in Set Theory’, *Annals of Pure and Applied Logic*, Vol. 164, 11: 1484-1523.

[Gödel 1931] Gödel, K. (1931), Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, *I. Monatshefte für Mathematik und Physik* 38: 173-98.

[Ganea 2009] Ganea, M. (2009), "Arithmetic on semigroups", *The Journal of Symbolic Logic*, vol. 74, pp.265–78.

[Gómez-Torrente 2004] Gómez-Torrente, M. (2004), 'The indefinability of truth in the "Wahrheitsbegriff" ', *Annals of Pure and Applied Logic* 126: 127–137. [?, Grover 1992] Grover, D., (1992), *A Prosentential Theory of Truth*, Princeton, NJ: Princeton University Press.

[Grover, et alii 1975] Grover, D., Camp, J., and Belnap, N., (1975), 'A Prosentential Theory of Truth', *Philosophical Studies*, 27: 73–125.

[Grzegorczyck 2005] Grzegorczyk, A. (2005), 'Undecidability without arithmetization', *Studia Logica*, 79, 305–313.

[Grzegorczyck&Zdanowski 2007] Grzegorczyk, A. and K. Zdanowski (2007), 'Undecidability and Concatenation', in Ehrenfeucht, A. et alii (eds.), *Andrzej Mostowski and Foundational Studies*, IOS, Amsterdam, pp. 72-91.

[Hájek&Pudlák 1993] Hájek, P., P. Pudlák (1993), *Metamathematics of First-Order Arithmetic*, Springer, Berlin.

[Halbach 1999] Halbach, V. (1999), 'Disquotationalism and Infinite Conjunctions', *Mind*, 108(429): 1-22.

[Halbach 2001a] Halbach, V. (2001a), 'How innocent is deflationism?', *Synthese* 126: 167-194.

[Halbach 2001b] Halbach, V. (2001b), 'Disquotational truth and analiticity', *The Journal of Symbolic Logic*, 66: 1959–1973.

[Halbach 2001c] Halbach, V. (2001c), *Deflationist Semantics*, habilitation thesis, Konstanz.

[Halbach 2009] Halbach, V. (2009), "Axiomatic Theories of Truth", *The Stanford Encyclopedia of Philosophy (Winter 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2009/entries/truth-axiomatic/>.

[Halbach 2011] Halbach, V. (2011), *Axiomatic Theories of Truth*, CUP.

[Halbach&Leigh 2013] Halbach, V. , G. Leigh, *String Theory*, unpublished draft.

[Halbach&Horsten 2014] Halbach, V. and L. Horsten (2014), 'Norms for reflexive theories of truth', in Achourioti, D., H. Galinon& J. Martinez (eds.) (forthcoming), *Unifying the Philosophy of Truth*, Springer.

[Heck 2009] Heck, R. (unpublished), *The Strength of Truth Theories*, unpublished manuscript.

[Hermes 1938] Hermes, H. (1938), *Semiotik. Eine Theorie der Zeichengestalten als Grundlage für Un- tersuchungen von formalisierten Sprachen, vol. 5 of Forschungen zur Logik und zur Grundlegung der exakten Wissenschaften*, Neue Folge, S. Hirzel, Leipzig.

[Hilbert 1950] Hilbert, D. and W. Ackermann (1950), *Principles of Mathematical Logic*, Chelsea Publishing Company, New York.

[Hilbert 1967a] Hilbert, D. (1967), "The foundation of mathematics", in Van Heijenoort (ed.) (1967), *From Frege to Gödel: a source book in mathematical logic*, Cambridge University Press (1967).

[Hilbert 1967b] Hilbert, D. (1967), "On the infinite", in Van Heijenoort (ed.) (1967), *From Frege to Gödel: a source book in mathematical logic*, Cambridge University Press (1967).

[Hodes 1984] Hodes, H. (1984), "Logicism and the ontological commitment of arithmetic", *Journal of Philosophy* 81: 123-149.

[Hodges 2008] Hodges, W. (2008), 'Tarski's Theory of Definitions', in [Patterson 2008, pp. 94-132].

[Hofweber 2000] Hofweber, T. (2000), 'Proof-theoretic reduction a a philosopher's tool', *Erkenntnis* 53: 127-146.

[Horsten 1995] Horsten, L. (1995), 'The Semantical Paradoxes, the Neutrality of Truth and the Neutrality of the Minimalist Theory of Truth', In: *The Many Problems of Realism (Studies in the General Philosophy of Science: Volume 3)*, P. Cortois (ed.), Tilburg: Tilburg University Press, 173-187.

[Horsten 2011] Horsten, L. (2011), *The Tarskian Turn: deflationism and axiomatic truth*, MIT Press, Cambridge MA.

[Horwich 1998] Horwich, P. (1998), *Truth*, 2nd ed., Blackwell.

[Jech 2003] Jech, T. (2003), *Set Theory: The Third Millennium Edition*, Springer, Berlin.

[Jensen&Karp 1971] Jensen, R. B. and C. Karp, 'Primitive Recursive Set Functions', in Scott D. S. (1971) (ed.), *Axiomatic Set Theory (Proceedings of the Symposia in Pure Mathematics, Vol. XIII, Part I, University of California, Los Angeles 1967)*, AMS, Providence.

[Joosten 2003] Joosten, J., 'Comparing $I\Sigma_1$ and PRA', in *Logic Preprint Series*, University of Utrecht.

[Kaye 1991] Kaye, R. (1991), *Models of Peano Arithmetic*, Oxford University Press.

[Kaye&Wong 2003] Kaye, R. and T. L. Wong, *On Interpretations of Arithmetic and Set Theory*, manuscript.

[Ketland 1999] Ketland, J. (1999), "Deflationism and Tarski's Paradise", *Mind*, 108(429), 69–94.

[Ketland 2011] Ketland, J. (2011), unpublished note.

[Kirby 2009] Kirby, L. (2009), 'Finitary Set Theory', in *Notre Dame Journal of Formal Logic* 50(3) 2009.

[Kirkham 1992] Kirkham, R. (1992), *Theories of Truth. A critical introduction*, MIT Press, Cambridge.

[Kotlarsky et alii 1981] Kotlarski, H., S. Krajewski and A. Lachlan (1981), 'Construction of satisfaction classes for nonstandard models', *Canadian Mathematical Bullettin* 24: 283-293.

[Kotlarski 1986] Kotlarski, H. (1986),'Bounded Induction and Satisfaction Classes', *Zeitschrift für Mathematische Logik*, 32, pp. 531Ð44.

[Lachlan 1981] Lachlan, A. (1981), 'Full satisfaction classes and recursive saturation', *Canadian Mathematical Bullettin* 24: 295-297.

[Lavine 1999] Lavine, S. (1999), *Skolem was wrong*, unpublished manuscript.

[Leeds 1978] Leeds, S. (1978), 'Theories of Reference and Truth', *Erkenntnis* 13: 111-127.

[Leigh 2013] Leigh, G. (2013), *Deflating Truth*, unpublished manuscript.

[Leigh&Nicolai 2013] Leigh, G. and C. Nicolai (2012), 'Axiomatic Truth, Syntax and Metatheoretic Reasoning', forthcoming in *The Review of Symbolic Logic*.

[Leitgeb 2007] Leitgeb, H. (2007), 'What theories of truth should be like (but cannot be)', *Philosophy Compass* 2,2: 276-290.

[Lewy 1947] Lewy, C. (1947), 'Truth and Significance', *Analysis* 8: 24-27.

[Lindström 2003] Lindström, P. (2003), *Aspects of Incompleteness*, ASL Lecture Notes in Logic.

[Mahn 1967] Mahn, F.-K., (1967), 'Zu den primitive-rekursiven Funktionen über einem Bereich endlicher Mengen, *Archiv für mathematische Logik und Grundlagenforschung*, 10: pp. 233–238.

[Montagna&Mancini] Montagna, F. and A. Mancini (1994),'A minimal predicative set theory', *Notre Dame Journal of Symbolic Logic* 35(2).

[Mancini&Zambella 2001] Mancini, A. and D. Zambella (2001), 'A note on recursive set theories', *Notre Dame Journal of Formal Logic* 42(2):109–115.

[Mancosu 1991] Mancosu, P. (1991), "Generalizing classical and effective model theory in theories of operations and classes", in *Annals of pure and applied logic*, 52, 3, pp. 249-308.

[Mancosu et alii 2009] Mancosu P., R. Zach, C. Badesa, (2009), "The development of mathematical logic from Russell to Tarski, 1900-1935", in L. Haaparanta (ed.), *The development of modern logic*, Oxford University Press, 2009.

[Manzano 1996] Manzano, M. (1996), *Extensions of first-order logic*, Cambridge University Press.

[Mathias 2006] Mathias, A. (2006), 'Weak systems of Gandy, Jensen and Devlin', in: J. Bagaria, S.Todorčvić (Eds.), *Set Theory: Centre de Recerca Matemàtica*, Barcelona 2003–4, Birkhäuser Verlag, Basel, 2006, pp. 149–224.

[McGee 1997] McGee, V. (1997), 'How we learn mathematical language', *The Philosophical Review*, 106, Vol. 1: 35-68.

[Murawski 1998] Murawski, R. (1998), 'Undefinability of Truth. The problem of priority: Tarski vs. Gödel', *History and Philosophy of Logic*, 19:3, 153-160.

[Montagna&Mancini 1994] Montagna, F. and A. Mancini, "A minimal predicative set theory", *Notre Dame Journal of Formal Logic*, 35, pp. 186-203.

[Niebergall 2011] Niebergall, K.G. (2011), 'Mereology', in Horsten L., Pettigrew, R. (eds.) (2011), *Continuum Companion to Philosophical Logic*, Continuum, London.

[Patterson 2008] Patterson, D. (ed.) (2008), *New Essays on Tarski and Philosophy*, Oxford University Press.

[Pedersen&Rossberg 2010] Pedersen N. and M. Rossberg (2010), 'Open-Endedness, Schemas and Ontological Commitment', in *Noûs* 44(2): 329-339.

[Pedersen&Wright 2013a] Pedersen N.J.L.L. and C. D. Wright (eds.) (2013), *Truth and Pluralism. Current Debates.*, Oxford University Press.

[Pedersen&Wright 2013b] Pedersen N.J.L.L. and C. D. Wright (eds.) (2013), "Pluralist Theories of Truth", *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL=<http://plato.stanford.edu/archives/spr2013/entries/truth-pluralist/>.

[Pettigrew 2009] Pettigrew, R. (2009), 'On interpretations of bounded arithmetic and bounded set theoryÕ, in *Notre Dame Journal of Formal Logic* 50(2):141-152.

[Previale 1994] Previale, F. , 'Induction and foundation in the theory of hereditarily finite sets', *Archive for Mathematical Logic* 33, pp. 213–41.

[Pohlers 2009] Pohlers, W. (2009), *Proof Theory. The First Step into Impredicativity.*, Springer.

[Prior 1971] Prior, A, (1971), *Objects of Thought*, Oxford: Clarendon Press.

[Pudlák 1983] Pudlák, P. (1983), 'Some prime elements in lattice of interpretability types', *Transactions of the American Mathematical Society* 280: pp. 255-75.

[Putnam 1978] Putnam, H. (1978), *Meaning and the Moral Sciences*, Routledge, London.

[Quine 1946] Quine, W. V. (1946), "Concatenation as a basis for arithmetic", *Journal of Symbolic Logic*, 11,4.

[Quine 1981] Quine, W. V. (1981), *Mathematical Logic*, (Revised Edition), Harvard University Press, Cambridge-London.

[Quine 1986] Quine, W.V. (1986), *Philosophy of Logic*, Harvard University Press.

[Quine 1990] , Quine, W.V. (1990), *The Pursuit of Truth*, Harvard University Press, Cambridge MA.

[Ramsey 1927] Ramsey, F. P. (1927), "Facts and Propositions",*Proceedings of the Aristotelian Society* supp. vol. 7:153Ð170.

[Rödding 1967] Rödding, D. (1967), 'Primitiv-rekursive Funktionen über einem Bereich endlicher MengM97en', in *Archiv für mathematische Logik und Grundlagenforschung* 10: pp. 13-29.

[Russell 2010] Russell, B. (2010), *The philosophy of Logical Atomism*, Routledge; originally appeared in *The Monist*, 1918.

[Shapiro 1998] Shapiro, S. (1998), "Truth and Proof: through thick and thin", *Journal of Philosophy*.

[Shapiro 2002] Shapiro, S. (2002), "Deflation and Conservation", in Halbach V. and L. Horsten (2002) (eds.), *Principles of Truth*, Dr. Hänsel-Hohenhausen, Frankfurt.

[Simpson 2009] Simpson, S. (2009), *Subsystems of Second Order Arithmetic*, 2nd ed., Cambridge University Press.

[Skolem 1919] Skolem, T. (1919), 'Untersuchungen über die Axiome des Klassenkalküls und über Produktations und Summationsprobleme, welche gewisse Klassen von Aussagen betreffen', *Videnskapsselskapets Skrifter, I. Matem.-naturv. klasse*, 3; reprinted in T.

Skolem, *Selected Works in Logic*, J. E. Fenstad (ed.), Oslo: Universitetforlaget, pp. 67Đ101.

[Smorynski 1977] Smorynski, C. (1977), 'The incompleteness theorems', in Barwise, J. (1977), *Handbook of Mathematical Logic*, North-Holland, Amsterdam.

[Soames 2010] Soames, S. (2010), *Philosophy of Language*, Princeton University Press, Princeton.

[Strawson 1949] Strawson, P. (1949), 'Truth', *Analysis* 9: 83-97.

[Švejdar 2009] Švejdar, V. (2009), ' On interpretability in the theory of concatenation', *Notre Dame Journal of Formal Logic* 50, pp. 87-95.

[Świerczkowski 2003] Świerczkowski, S. (2003), Finite Sets and Gödel's Incompleteness Theorems, in *Dissertationes Mathematicae* 422.

[Tait 1981] Tait, W. (1981), 'Finitism', in *Journal of Philosophy* 78, pp. 524-546.

[Tarski 1939] Tarski, A. (1939), 'New investigations on the completeness of deductive theories', *Journal of Symbolic Logic* 4(1939): p. 176.

[Tarski 1956a] Tarski, A. (1956a), "'The concept of truth in formalized languages", in Woodger, H. J. (trans., ed.), *Logic, Semantic, Metamathematics: papers of Alfred Tarski from 1922-1938*, Oxford, Clarendon Press, pp. 152-278.

[Tarski 1956b] Tarski, A. (1956b), "Some observations on the concepts of ω-consistency and ω-completeness", in *Logic, Semantic, Metamathematics: papers of Alfred Tarski from 1922-1938*, Oxford, Clarendon Press, pp. 279-295.

[Tarski 1944] Tarski, A. (1944), 'The semantic conception of truth', *Philosophical and Phenomenological Research*, 4, 341-375.

[Tarski et alii 1953] Tarski, A., A. Mostowski, and R. M. Robinson (1953), *Undecidable Theories*, North Holland, Amsterdam.

[Tarski&Givant 1987] Givant, S. and A. Tarski (1987), *A formalization of set theory without variables*, vol. 41 of *American Mathematical Society Colloquium Publications*, AMS, Providence.

[Troelstra&Schwichtenberg 2000] Troelstra, A. S. and H. Schwichtenberg (2000), *Basic Proof Theory*, number 43 in Cambridge tracts in theoretical computer science, Cambridge University Press, Cambridge, second edition.

[Van Dalen 1980] Van Dalen, D. (1980), *Logic and Structure*, Springer.

[Varzi 2012] Varzi, A. (2012), 'Mereology', *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/mereology/>.

[Visser 2008a] Visser, A. (2008), 'Growing commas. A study of sequentiality and concatenation', *Logic Group Preprint Series* 257.

[Visser 2008b] Visser, A. (2008), 'Pairs, Sets and Sequences in First-Order Theories', *Archive for Mathematical Logic* 47: pp. 299-326.

[Visser 2008c] Visser, A. (2008), 'Categories of theories and interpretations', *Logic in Tehran*, Lect. Notes Log., vol. 26, Assoc. Symbol. Logic, La Jolla, CA, 2006, pp. 284–341. MR MR2262326 (2007j:03083)

[Visser 2011] Visser, A. (2011), 'Interpretations', Lecture Notes , *Summer School on Set Theory and HIgher Order Logic*, London, August 2011.

[Von Neumann 1966] Von Neumann, J. (1966), *Theory of Self-Reproducing Automata*, A. W. Burks (ed.), University of Illinois Press, Urbana.

[Wang 1981] Wang, H. (1981),'Some facts about Kurt Gödel', *The Journal of Symbolic Logic* 46: 653-659.

[Wang 1987] Wang, H. (1987), *Reflection on Kurt Gödel*, MIT Press.

[Wetzel 1989] Wetzel, L. (1989), "Expressions vs Numbers", *Philosophical Topics*, XVII, 2.

[Wetzel 2009] Wetzel, L. (2009), *Types and Tokens: on Abstract Objects*, MIT Press.

[Williams 1976] Williams, C. (1976), *What is truth?*, Cambridge University Press, Cambridge.

[Williams, M. 1986] Williams, M. (1986), 'Do We (Epistemologists) Need a Theory of Truth?', *Philosophical Topics* 14:223–242.

[Wolenski 1989] Wolenski, J. (1989), *Logic and Philosophy in the Lvov-Warsaw School*, Dordrecht: Kluwer.