Uniwersytet Warszawski
Wydział Filozofii i Socjologii

Bartosz Wcisło
Nr albumu: 276697

Understanding the Strength of the Compositional Truth

Praca doktorska
na kierunku FILOZOFIA

Praca wykonana pod kierunkiem
dra hab. Cezarego Cieślińskiego

Październik 2017

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Abstract

We investigate the properties of compositional truth theories and try to understand their strength. We are primarily concerned with two notions of strength: proof-theoretic and model-theoretic. The first notion measures the consequences of a given theory in the language of its base theory. Our main result states that the theory of compositional truth with bounded induction is not conservative over Peano arithmetic. Model-theoretic strength measures how much restriction is put on models of the base theory which are expandable to a model of a given truth theory. One of our other main theorems shows that every model of Peano arithmetic which is expandable to a model of the compositional theory of truth also allows an expansion to a model with a truth predicate satisfying full induction and uniform Tarski biconditionals. We also present our findings concerning model-theoretic strength of truth theories with compositional axioms based on strong Kleene logic rather than classical logic. The results presented in our thesis have been obtained in joint work with Cezary Cieśliński and Mateusz Łełyk.

Słowa kluczowe

truth theories, compositional truth predicate, conservativity, proof-theoretic strength, model-theoretic strength, models of truth theories, CT_0

Dziedzina pracy (kody wg programu Socrates-Erasmus)

8.1 Filozofia

Tytuł pracy w języku polskim

Siła kompozycyjnych teorii prawdy

Contents

Introduction

This thesis concerns formal truth theories. On a very general level, it is a study of the notion of truth and certain philosophical questions concerning this notion using formal methods.

Overview of the research

Theories of truth as considered in this thesis are modelled in the following fashion: we fix a first-order theory B (a **base theory**) which models at least our theory of syntax and we add a unary predicate $T(x)$ with the intended reading "x is a true sentence" along with the axioms governing this predicate. Thus we obtain a **truth theory** Th. We then ask all sorts of questions about the obtained theory and its relations with the theory B.

Note that this is very different from the original approach of Tarski, who defined for a given language $\mathscr{L}$ what it means for a sentence to be true in a structure interpreting symbols from $\mathscr{L}$. We definitely do not want to *define* what a true sentence is. We *postulate* that there is some property with reasonably truth-like behaviour and try to understand how the presence of that predicate changes the properties of our theory. The properties that we postulate are typically some restrictions of the most obvious and natural ones, since the natural ones have a rather unpleasant tendency to be inconsistent.

Generally, we expect that adding a truth-like predicate to a base theory B will yield a theory stronger than B. There are many specific incarnations of this general phenomenon in diverse areas of logic. The most basic result, and one of the most famous along these lines, is Tarski's Theorem which states that no theory which has a modicum of arithmetic can define its own truth predicate. Thus, we see the first possible sense in which a theory of truth Th can be stronger than B: it expresses a notion which is not definable in the base theory B alone.

There are other possible interpretations of the notion of strength which can help formalise the intuition that, when added to the base theory B, the truth predicate yields a theory which is substantially stronger. The two which we focus on in this thesis are as follows:

- Theory of truth Th proves more sentences *in the language of* B than B itself.
- Theory of truth Th puts more restriction on models than B itself: i.e. not in every model of B, can we find an interpretation for the truth predicate.

In the first condition, we demand that the theory of truth proves some new theorems *in the language of* B. Obviously, if we add a new predicate to a language, then the resulting theory will inevitably prove some new theorems, e.g., propositional tautologies in the language of the expanded theory. To avoid this kind of counterexamples, we make the explicit demand that a sentence counts as a nontrivial theorem only if it *could*, in principle, be proved in B alone. Adding a truth predicate allows us to draw new conclusions about matters which are not explicitly related to truth. In such a case, let us say that B is **syntactically stronger** than Th or **syntactically nonconservative**. Obviously, it is syntactically conservative otherwise.

The second condition is very similar in spirit. We demand that there are models of B in which there is no interpretation for the truth predicate. This means that the presence of the truth predicate excludes some possible structures which *could be* models of B. Thus, the truth theory is more restrictive than the base theory when it comes to models. In such a case, let us say that Th is **semantically stronger** than Th or, alternatively, that Th is **semantically nonconservative** over B. Again, Th is semantically conservative if it is not semantically nonconservative. These two notions are related in the following way: if Th is syntactically stronger than B, then it is also semantically stronger which is an easy consequence of the completeness theorem for first-order logic. In general, this implication does not reverse.

As will be explained in the thesis, truth theories may or may not be stronger than the base theory, depending on the axioms for the truth predicate. As stated, this is not at all surprising, because there are obviously some systems of axioms which do not even make the predicate $T(x)$ look like the formalisation of the notion of truth, e.g. the one given by the axiom $\forall x\, T(x)$. However, as we will see, we can find many systems of axioms which yield the predicate T to be undoubtedly truth-like and yet do not make the truth theory stronger than B.

The over-arching problem with which this thesis deals is as follows: What is the dividing line between strong and weak truth theories? What are the principles that cause weak theories to become strong? We are focused on isolating the borderline cases where the addition of innocuous axioms to weak theories of truth makes them strong. Of course, this goal is so broad that it has to be narrowed down in some way; indeed, we will restrict our attention to a few particularly natural cases.

Thus, to repeat ourselves: our main goal is to understand where the transition point between weak and strong theories of truth is located. Hopefully and in the long run, this can let us understand why certain principles make the theory of truth stronger than its base theory.

Philosophical background

In this section, we shall briefly present the philosophical background from which our research originates and try to discuss the relevance of the results which we may obtain within our approach in this philosophical context.

The study of the strength of truth as presented in this thesis originates in the study

on the deflationary theory of truth. **Deflationism** claims that truth predicate does not have any other role or meaning other than that it satisfies so called **Tarski's biconditionals**, i.e., sentences of the form:

$\ulcorner\phi\urcorner$ is true if and only if ϕ,

where $\ulcorner\phi\urcorner$ is a name for the sentence ϕ. To quote an actual deflationist:[1]

> It can be argued that such [Tarski's] biconditionals are *epistemologically fundamental*:—we do not arrive at them, or seek to justify our acceptance of them, on the basis of anything more obvious or more immediately known. It can be argued, in addition, that our underived inclination to accept these biconditionals is the source of *everything else* we do with the truth predicate.

Thus, the deflationist is committed to the following **positive claim**:

- Tarski's biconditionals fully explain the behaviour of the truth predicate as present in the natural language.

Deflationists typically claim that the right theory of truth should consist in Tarski's biconditionals. This position is not defended by *all* adherents of deflationism. For example, in [Horsten, 2009], Horsten presents the stance that a theory of truth in which compositional clauses for the truth predicate are reformulated as rules (e.g., enabling us to deduce that $\phi \wedge \psi$ is true from the assumption that ϕ is true and ψ is true) is also acceptable for the deflationist.

Thus the positive claim of deflationists is that all properties of the truth predicate present in the natural language can be explained using some very minimal part of what seems to be the properties of the truth predicate. The main thesis of the deflationary theory of truth may be rephrased as follows:

- To explain the behaviour of the truth predicate we *do not have to* postulate any further properties of this notion which do not already consist in satisfying Tarski's biconditionals.

In particular, we do not have to postulate that the truth predicate ascribes to sentences some property which can be *further explained*, like verifiability or correspondence. Deflationists claim that there is nothing really more to say about our truth talk other than that the truth predicate satisfies Tarski's biconditionals. In such a way, we avoid any obligation to explain what the semantic notions actually *mean*. A large group of philosophical problems concerning semantics thus disappears once we accept this solution.

The claim presented above is not really different from the positive claim. In the first place, we have explicitly written it down to contrast it with another thesis which has been associated with deflationism, the **negative claim**:

- Truth is an insubstantial notion.

[1] [Horwich, 2001], p. 149.

The word "insubstantial" is notoriously imprecise in this context. That the notion of truth is insubstantial roughly means that it has no interesting metaphysical or epistemological content. One of the possible explications of the negative claim is to demand that the deflationary theory of truth be conservative over the base theory. This is intended to model that a correct truth theory does not allow us to infer any new insights about the non-semantic part of the world. The explication in question may seem somewhat hasty, as the thesis that truth is an "insubstantial" notion and the thesis that truth theory should be conservative do not seem very strongly related. In fact, whether such an explication is satisfactory has been questioned in the literature.[2] On the other hand, Hartry Field, who is one of the prime proponents of deflationism, seems to agree that adding purely truth-theoretic axioms should form a conservative extension of the base theory (by which he also seems to concede to some version of the negative claim).[3] Note that we do not *equate* "insubstantiality" with "conservativity." The explication in question only means that the negative claim of deflationism should, in particular, *entail* that truth theory is conservative over the base theory.

We refrain from answering the question on the correctness of the conservativeness explication of the negative claim of deflationism. We personally lean towards the opinion that there are a number of possible formal explications to this thesis which can be effectively viewed as expressing genuinely different philosophical stances sharing some common intuitions. Thus, the considerations in our thesis are simply most relevant to these versions of deflationism which claim that the theory of truth should be conservative over the base theory of syntax.

The deflationary theory of truth has been famously criticised by Jeffrey Ketland in [Ketland, 1999] and Stewart Shapiro in [Shapiro, 1998]. Their arguments are based on similar motives, but are substantially different nonetheless. Ketland's argument may be summarised (roughly) as follows:

- A theory of truth whose axioms state that the truth predicate is compositional with respect to the base language (and satisfies the full induction scheme for the sentences containing truth predicate) proves that whatever is provable in our base theory is true. Therefore, it proves the consistency of the base theory.
- We have good reason to accept the compositional axioms and the full induction scheme.
- Any adequate theory of truth proves that whatever is provable in our base theory is true. In particular, it proves the consistency of the base theory.[4]

[2] See [Horsten, 2009] and [Cieśliński, 2015a] where it is claimed that non-conservative theories of truth may be acceptable from the deflationist's point of view.

[3] See [Field, 1999], especially pp.536–537 where he agrees that if induction axioms "depend on the nature of truth," then under some mild additional assumptions, truth cannot be deflationary because the theory of inductive compositional truth is not conservative over the base theory.

[4] It is not entirely clear to us whether this requirement in [Ketland, 1999] is based on the more basic premise that truth satisfies compositional clauses and induction (which seems to be implied by remarks on p. 91), or is simply an independent assumption (as suggested by the last paragraph on p. 90). This is not very relevant for our further considerations.

- A theory of truth whose axioms are based on Tarski's biconditionals for the sentences of the base language does not prove the consistency of the base theory.
- Therefore, the deflationist should not accept Tarski's biconditionals as the sole axioms for the truth predicate.

The argument of Shapiro revolves around very similar considerations, but, in a way, goes the other way round. It may be outlined as follows:

- Since deflationists claim that truth is an insubstantial notion, truth theory should not allow us to prove any new sentences in the language of the base theory.
- Adding to our theory a compositional truth predicate (which satisfies the full induction scheme) indeed allows us to prove new sentences in the language of base theory.
- We have good reason to accept the compositional axioms and the full induction scheme.
- Therefore the deflationist should not claim that truth is insubstantial.

The third point is actually implicit in Shapiro's argument, since he simply does not consider an option that a deflationist might be willing to reject outright, or just doubt the compositionality of the truth predicate.

Notice that, in a way, the two arguments are exactly opposite. Ketland claims that any adequate truth theory *certainly* proves more sentences in the language of the base theory than the base theory itself and therefore cannot be based on Tarski's biconditionals, whereas Shapiro claims that *since* truth theory is compositional and inductive, it cannot be syntactically conservative over the base theory. It seems that Ketland's argument is mainly directed against deflationists who claim that the functioning of the truth predicate is fully encapsulated in Tarski's biconditionals and Shapiro's argument is directed against deflationists who claim that truth is an insubstantial notion.

The arguments sketched above have launched the research on conservativity of formal truth theories in the context of the deflationism debate. Note that both the arguments of Shapiro and of Ketland are aimed to provide a firm grounding for this discussion. We refrain from presenting its further stages here. What we will do instead is to try to assess what is the potential importance of any conservativity results for the debate outlined based on the objections of Shapiro and Ketland.

The first observation is that neither Shapiro's nor Ketland's argument aims to show that the negative and the positive claim of the deflationary theory of truth are inconsistent (which is simply not the case). In Shapiro's argument, we assume that the correct theory of truth contains compositional axioms and induction while, in Ketland's argument, we require that using the truth predicate one can show that theorems of the base theory are true (in particular, we require that it is nonconservative).

Hence, it seems that the most natural response for the deflationist should be simply to embrace Tarski's biconditionals as the only axioms for the truth predicate. This

would exactly result in obtaining a conservative theory of truth. A deflationist could simply agree that based on Tarski's biconditionals we really cannot infer that truth is compositional and this is precisely the argument for the claim that we have no justification for stronger concepts of truth. As those two basic intuitions of deflationism are actually mutually consistent, the arguments against deflationism must directly attack at least one of them (that is, assuming that there are no other claims whose position to the deflationary theory of truth is as central as the positive and the negative claim as described above).

In our thesis, we will present a number of results to the effect that a certain theory of truth is conservative or that it is not conservative. It would appear that they might potentially have a twofold use in the conservativeness debate; one related to Shapiro's argument and the other to Ketland's.

First of all, we will see in our thesis that a number of seemingly very innocent-looking truth-theoretic principles are not in fact conservative over PA. This finding seems quite relevant to the debate on the deflationary theory of truth. In his argument, Shapiro crucially uses the induction axioms for the language containing truth predicate. As we shall see, a truth theory which comprises only Peano arithmetic and the compositional clauses for the truth predicate is actually conservative over PA. This has led to the following response of Hartry Field:[5] Shapiro's conclusion that truth theory is not syntactially conservative is based not on purely truth-theoretic considerations, but also on induction principles for the truth predicate. According to Field, these new induction axioms are really new assumptions about the structure of natural numbers which could not be spelled out before we introduced a new predicate. Unlike compositional clauses, these axioms cannot be justified basing purely on considerations about the concept of truth. They really express our insights about numbers.

Here it is clear that some findings about non-conservativeness of truth-theoretic principles can be directly relevant to the debate on deflationism. If we can locate some theories which apparently employ only plausible purely truth-theoretic principles, then once we manage to show that these theories are not conservative over PA, a deflationist will be forced either to abandon the conservativeness claim or to deny that the truth predicate, in fact, obeys these strong principles. Some such particularly simple properties have been pointed out by Cieśliński.[6]

Unfortunately, the argument of Cieśliński was (at least partly) based on some older well-known theorem whose proof was shown to contain a gap.[7] In our thesis, we provide a different proof of that theorem.

Let us consider whether investigation on the strength of various truth theories may be relevant to the conservativeness argument in Ketland's version. By his account, any adequate theory of truth must be capable of proving, for instance, the consistency of

[5]See [Field, 1999].

[6]See [Cieśliński, 2010a], Theorems 3 and 4.

[7]Namely, Theorem 4 of [Cieśliński, 2010a] states that from the principles "truth is compositional" and "whatever is provable from true premises in propositional logic is true" induction for Δ_0-formulae containing the truth predicate may be derived which was previously claimed to be non-conservative over PA.

its base theory. Thus, it cannot be conservative. If we accept this requirement, then the theories which turn out to be conservative are by the same token shown to be inadequate. Therefore, considerations about the strength of truth theories could in principle show that, in fact, certain possible deflationist proposals of the correct set of axioms for the truth predicate are untenable.

In effect, the results on strength of truth theories seem to have two applications to Ketland's argument. The first one precisely resembles the case of Shapiro's argument: if we find out that some seemingly innocent theories of truth are in fact non-conservative over the base theory, this is a further support for Ketland's claim that adequate theories of truth should not be conservative. This support is of limited relevance. Unlike Shapiro, Ketland does not find non-conservativeness of truth theories *in itself* problematic for deflationists. What is problematic is that the usual disquotational truth theories *are conservative* over the base theory.

Research on axiomatic theories of truth could also be relevant for the discussion of Ketland's argument in a different way. If we find some theories of truth which are, in some sense, based on disquotational scheme but prove all the consequences that Ketland requires from an adequate theory of truth, then this could provide a possible way of defending deflationism against the discussed argument. Such strong theories essentially based on a disquotational scheme, have already been investigated in the literature, see e.g. [Halbach, 2009] or [Horsten and Leigh, 2017].[8] However, this is outside the scope of our thesis which deals mainly with compositional truth theories.

It seems that the discussion of both Shapiro's and Ketland's arguments can be influenced by formal results on conservativity of truth theories. Having said that, we want to stress that there is a serious limitation to any possible use of the results on strength of formal theories of truth in the debate on conservativeness argument. Namely, as we have already remarked, two claims of deflationism (positive and negative) are in fact mutually consistent. If the deflationist is ready to say that really the only valid truth theoretic principle are Tarski's biconditionals and *prima facie* we have no reason to accept compositional clauses, *therefore* truth is a insubstantial notion which does not allow us to draw any nontrivial conclusions about the surrounding world, then this position is internally coherent.

Whether or not the deflationist's claims are immune to the conservativeness argument (in any of the discussed variants) presumably depends on the main source of justification for those claims. The first possible approach is present in the quoted fragment from Horwich. Namely, the deflationist has some good reasons to believe that our theory of truth consists fully in Tarski's biconditionals (or some form thereof). All other claims are secondary to this one. Basically, it is a claim that *contrary to what may appear* what we really assume when we use the truth predicate is that it satisfies Tarski's biconditionals. It is similar in spirit, say, to Hume's claim that contrary to what may appear, the only thing which we mean when we speak about causality reduces to facts

[8] In [Halbach, 2009], an untyped truth predicate is considered, i.e., the truth predicate satisfying Tarski's biconditionals for some sentences in which that very predicate occur. In [Horsten and Leigh, 2017], disquotational axioms are augmented with reflection principles.

about time-ordering of events.

In this case, a serious objection to deflationism can be formulated on the basis of Ketland's argument. The objection is very simple: truth theory comprising compositional axioms and full induction is not conservative over the base theory. Disquotational truth theory is conservative provided that the truth predicate is typed (i.e., that the disquotational axioms only hold for sentences from the language of the base theory). Hence compositional axioms cannot be derived from disquotational axioms in the case of untyped truth theories.[9] If the deflationist claims that the primary meaning of the truth predicate is encapsulated in Tarski's biconditionals and the rest of our practices concerning the notion of truth can be explained on the basis of these axioms, then a problem occurs: In what sense can compositionality of truth predicate be explained on the basis of disquotation scheme if compositional axioms *cannot* be derived from disquotational ones?

Note that here we are only concerned with a disquotationalist who primarily claims that, in fact, we do not need anything apart from disquotational scheme to explain the functioning of the notion of truth. The fact that we apparently *need* some additional axioms to describe how our natural notion of truth functions appears to be a rather direct argument against this version of deflationism.

The deflationist could in fact claim that the *meaning* of the truth predicate entails only Tarski's clauses, but still make a *quasi-empirical* hypothesis that it actually behaves compositionally. As a matter of fact, any concrete instance of this general phenomenon could be even proved from the deflationist's axioms. For instance, for any particular arithmetical sentence ϕ, it can be proved by using only Tarski's biconditionals that ϕ is true if and only if the negation of ϕ is not true. What is beyond the scope of Tarski's biconditionals is to prove the general fact that the truth predicate behaves compositionally for all arithmetical sentences. As we have already remarked, this is typically also true in the case of untyped truth theories even though the argument based on the conservativity need not carry over to this case.

Another possible strategy of defending the view that all properties of the truth predicate are encapsulated in Tarski's scheme is to work with stronger logic (e.g. second-order or infinitary logic) or, generally, to employ some additional means of reasoning. For instance, to assume that we may close off our disquotational theory of truth under the reflection scheme. Such additional means may allow us to derive compositional clauses from the disquotational scheme.[10] (In fact, they can yield much more than that.) However, we shall not explore such strategies in our thesis.

Above, we have discussed the implications of conservativity arguments for the variant of deflationism in which the primary claim is that the functioning of the notion of truth is fully described by Tarski's scheme and the insubstantiality claim is secondary to the positive claim. However, Shapiro's conservativeness argument seems directed against the negative claim of the deflationary theory of truth. Therefore, it appears

[9] Typically, they cannot also be derived in the untyped case (where disquotational axioms are included for sentences containing the truth predicate), for quite different reasons, but let us focus on the simpler case.

[10] The case of the reflection scheme has been investigated in [Horsten and Leigh, 2017].

reasonable to at least consider briefly what strategies of refuting that argument may exist for deflationists whose *core* claim is that truth theory is conservative over the base theory.

The technical crux of Shapiro's argument is the observation that the truth theory cannot at the same time be conservative (which he takes to be a consequence of being "insubstantial") and employ induction along with the compositional clauses. One position that this argument cannot affect is the one deeming it more fundamental that truth be conservative than it enjoy all the natural properties we believe that it should enjoy. If the deflationist sticks to the claim that truth theory is conservative over the base theory, then Shapiro's conservativeness argument can really *at best* show that an adherent of deflationism should be ready to accept that either compositional clauses or the induction axioms for the truth predicate may fail. Nonetheless, Shapiro's argument still cannot really threaten a deflationary theory of truth which *builds on* the conservativity claim. On the other hand, Ketland's argument is not even aimed against the negative claim. He either assumes that truth theory is not conservative or deduces it from compositionality and induction, like Shapiro. Either way, the latter argument is no more problematic for the deflationist than the former one.

The position which places more trust in the negative claim than in the compositional clauses and induction is purely hypothetical. However, it seems that the very basic motivation of deflationism is not the quasi-empirical observation about the functioning of the truth predicate in the natural language, but rather the suspicion that there is something inherently wrong with the correspondence theory of truth. More generally, it is a suspicion that truth predicate is some kind of logical device which does not express a genuine property of sentences.

The deflationist does not necessarily discover that truth is insubstantial *on the basis* of the fact that this concept is encapsulated in Tarski's biconditionals. It seems that, at least in some cases, the positive claim is really a means of accounting for the functioning of the truth predicate in the natural language once we think that it does not function like any other normal predicate. In the above discussion, we have called the position that a truth predicate is not substantial "the negative claim of deflationism" and we (tentatively) agreed to explicate it as the conservativeness claim. We can then imagine that someone would be more willing to deny that truth is compositional or that it satisfies induction than to withdraw the negative claim of deflationism.

As we have already written, the version of deflationism sketched out above is hypothetical, being that we are not aware of its actual, explicit proponents. However, let us briefly argue that the claim that addition of compositional truth predicate may yield false consequences in the base language (or even may be inconsistent) is not *obviously* wrong.

One possible objection against this claim is that, as present in natural language, the truth predicate apparently enjoys full induction and satisfies the compositional clauses. The answer could be simply that the users of natural language can sometimes introduce predicates whose theory is not even consistent. As present in the natural language, the naïve truth predicate seems to be precisely an example of this phenomenon. Arguably, this predicate is naïvely assumed to satisfy all Tarski's biconditionals, even

untyped, thus making the truth predicate applicable to the sentences which contain that very predicate. In classical logic and with a modicum of arithmetic, this theory is simply inconsistent. Hence, the mere fact that our linguistic practices seem to include some axioms is not a good argument that these axioms hold.

Another apparent problem with the stance which expresses a preference for the negative claim over compositionality and induction for the truth predicate is that the standard model of arithmetic can be expanded to a model which satisfies these natural properties. This theory of truth is not conservative over PA which we accepted as our base theory. Hence, we know that some nonconservative theories of truth are in fact consistent. The response is that PA has been only chosen as the base theory for technical reasons and this argument does not generalise to all base theories. In particular, if we choose as our base theory the strongest theory that we actually accept then, by Tarski's theorem, we are no longer in a position to define the class of true sentences without resorting to the truth predicate, the same whose properties are being questioned. Thus, if one is sceptical about whether adjoining a compositional inductive truth predicate to our language could yield wrong consequences in the language of the base theory, then we cannot typically resort to an analogue of the construction of the arithmetical truth predicate for the standard model of arithmetic because we cannot guarantee that there is any analogue.

Overview of the thesis

In Chapter 1, we present all the necessary formal tools. We introduce basic notions of formal arithmetic which will serve as our base theory and the main truth theories considered in this thesis, as well as some basic technical lemmas which do not fit perfectly with any particular further chapter.

Chapter 2 deals with syntactic conservativeness of truth theories. In this chapter, we present a proof that CT_0—the compositional theory of truth with Δ_0-induction for formulae containing truth predicate—is not conservative over Peano Arithmetic.

In Chapter 3, we consider the semantic conservativeness of theories of truth. This chapter contains a proof that any model of PA expandable to a model of the compositional truth theory CT^- is also expandable to a model of UTB, i.e. the theory axiomatised by uniform Tarski biconditionals along with full induction. This fact implies Lachlan's theorem that for any model of CT^-, its restriction to arithmetical language is recursively saturated.

Chapter 4 deals with models of extensions of PT^-. This is a theory of truth whose compositional axioms are modelled after partial logic (Strong Kleene's Logic) rather than classical logic. We prove that PT^- extended with axioms of internal induction for total formulae is not semantically conservative over PA (contrary to what has been previously claimed in the literature). We also investigate PT^- extended with the unrestricted internal induction axiom and show that any model of this theory is also expandable to a model of UTB. In particular, every such model has to be recursively saturated.

In Chapter 5, we present the summary of the results. We also briefly discuss other notions of strength upon which we did not focus in the main part of the thesis.

The afore-mentioned results from Chapters 2 and 3 - and the first of the results from Chapter 4 - have been obtained by the author in cooperation with Mateusz Łełyk. The second of the mentioned results from Chapter 4 has been obtained in cooperation with Cezary Cieśliński and Mateusz Łełyk.

Chapter 1

Preliminaries

Truth theories are obtained by adding axioms governing a truth predicate to a given base theory, which we conceive a suitable candidate for a formalisation of our knowledge of the "extrasemantic" world, most notably of our theory of syntax. Therefore we have to describe our base theory—Peano Arithmetic and various axioms governing the truth predicate, which we will investigate further in our thesis.

1.1. Arithmetic

In this section, we provide the basic information concerning our base theory and its handling of syntax. All the facts mentioned in this section are utterly standard. In case of any doubts, the reader is referred to [Hájek and Pudlák, 1993] (especially Preliminaries, Chapters 1 and 3) and [Kaye, 1991], Chapters 1–5 and 9.

1.1.1. Peano Arithmetic

Let us begin with introducing our base theory—Peano Arithmetic. It is intended to describe the structure of natural numbers with the operations of addition, multiplication and the successor function.

Definition 1. By **Robinson Arithmetic** (Q) we mean the theory formulated in the language with no relation symbols, one constant 0, one unary function symbol S and two binary function symbols $+, \cdot$ whose axioms are universally quantified versions of the following clauses:

1. $x \neq y \rightarrow S(x) \neq S(y)$.
2. $S(x) \neq 0$.
3. $x + 0 = x$.
4. $x + S(y) = S(x + y)$.
5. $x \cdot 0 = 0$.

6. $x \cdot S(y) = x \cdot y + x$.

7. $x \neq 0 \to \exists y\ S(y) = x$.

We call the language of Q the **arithmetical language** and denote it with $\mathscr{L}_{\mathrm{PA}}$.

Sometimes we need to refer in our theory to a specific number. For an arbitrary $n \in \omega$ we define

$$\underline{n} = \underbrace{S \dots S}_{n \text{ times}} 0.$$

Thus $\underline{n}$ is a term obtained by preceding the symbol "0" by a sequence of n successor symbols "S". Terms of this form are called **numerals**. By convention, $\underline{0}$ is exactly the same term as 0 (it is the symbol "0" preceded by a sequence of symbols "S" of length 0, i.e. the empty sequence).

Definition 2. By **Peano Arithmetic** (PA) we mean the theory extending Robinson Arithmetic Q with the following axiom scheme:

$$\forall x_1, \dots, x_n \Big(\forall y \Big(\phi(y) \to \phi(S(y)) \Big) \longrightarrow \Big(\phi(0) \to \forall y\ \phi(y) \Big) \Big).$$

The scheme introduced above is called the **induction scheme**. It is the hallmark of PA. Basically, Robinson's Arithmetic axioms give inductive definitions of the successor, addition and multiplication functions in natural numbers $\mathbb{N}$. The inductive definitions themselves have little sense unless coupled with the induction axioms, which states that a property which holds of 0 and is preserved in passing to the successor of a given number holds of every number.

The set of natural numbers $\mathbb{N}$ together with the natural successor function S and binary functions $+$ of addition and $\cdot$ of multiplication form the canonical model $(\mathbb{N}, 0, S, +, \cdot)$ of Peano Arithmetic. It is called the **standard model**. We will denote it simply with $\mathbb{N}$. In Section 1.3, we will discuss models of PA in more depth.

It may seem awkward that we have chosen arithmetic as a base theory, since the latter is intended to capture the whole syntax. However, as we shall shortly see, PA is more than enough to this end. The technical crux is contained in the following remarkable theorem:

Theorem 3. *There exists a formula* $\exp(x, y) \in \mathscr{L}_{\mathrm{PA}}$ *such that* PA *proves the universal closure of the following formulae:*

1. $\exp(0, 1)$.

2. $\exp(S(x), y) \equiv \exists z \big(\exp(x, z) \wedge y = 2 \cdot z \big)$.

3. $\forall x \exists y\ \ \exp(x, y)$.

Intuitively, $\exp(x, y)$ holds only if $y = 2^x$. Therefore we can rewrite somewhat unintuitive clauses of the above theorem in the following way:

1. $2^0 = 1$.
2. $2^{S(x)} = 2 \cdot 2^x$.
3. $\forall x \exists y \ \ 2^x = y$.

In other words, there is an arithmetical formula satisfying the inductive clauses for the exponentiation such that the function given by this formula is provably total.

Convention 4. We will see a number of formulae defining functions similarly to $\exp(x, y)$. We will be writing these formulae in the functional notation. I.e., if theories which we consider prove that for all x there exists a unique y such that $\phi(x, y)$, we will generally write "$\phi(x) = y$" or even use the expression $\phi(x)$ as if it were a term. For such formulae ϕ, expressions of the form

$$\Psi(\phi(x))$$

should be understood as abbreviations for

$$\exists y \, \big(\phi(x, y) \wedge \Psi(y)\big).$$

1.1.2. Coding

The remarkable fact that PA defines exponentiation allows us to define syntactic notions and more generally, to recover a good part of set theoretic notions.

Definition 5. We define the arithmetical formula $x \in y$ as:

$$\exists a, b \big(a \neq 0 \wedge y = 2^x a + b \wedge b < 2^x\big),$$

where $u < v$ is an abbreviation for $\exists w \, \big(w \neq 0 \wedge u + w = v\big)$.

Intuitively, $x \in y$ means that x-th bit in the binary expansion of y is 1 rather than 0. The fact that we can define total exponentiation function together with induction axioms enables us to prove virtually all important basic properties of binary expansions. Thus we may view every number y as the finite set of all the numbers x such that x-th digit in binary expansion of y is 1. All these numbers may in turn be also viewed as sets and so on. This allows us to recover set theoretic notions, such as the **ordered pair** (x, y) defined as:

$$\{x, \{x, y\}\}.$$

The definition of ordered pair allows us to define the product of sets $A \times B$ as the set of ordered pairs (a, b) with $a \in A$ and $b \in B$. Thus we may define in PA relations as subsets of products of sets or functions as relations satisfying the usual conditions. In general, we may speak in PA of all finite set-theoretic objects: finite relations, graphs, trees, groups etc. In particular, if we identify symbols of $\mathscr{L}_{\mathrm{PA}}$ together with purely logical symbols, like connectives, quantifiers, brackets or variables with some fixed arbitrarily chosen numbers, we may speak in PA about strings of characters from the language of PA. Here, by a **string** we mean simply a finite sequence, i.e. a function whose domain

is some initial segment of natural numbers. A function a whose domain is $\{1, \ldots, n\}$ with values $a(i) = a_i$ will be sometimes denoted with

$$\langle a_1, \ldots, a_n \rangle.$$

Following the standard terminology, we call the values of a sequence its **terms** (not to be confused with a syntactic notion of term which we will shortly introduce). We call the cardinality of the domain of a sequence its **length**. We denote the length of a sequence a with $\mathrm{lh}(a)$. By convention, the (number identified with the) empty set is also a sequence and its length is 0. If a,b are two sequences of lengths n,m, respectively, then by **concatenation** of a with b we mean the unique sequence of length $n+m$, whose first n terms are $a(1), \ldots, a(n)$ and the next m terms are $b(1), \ldots, b(m)$. We denote the concatenation of a and b with

$$a \frown b.$$

We can speak within arithmetic about the arithmetical language itself. First, we simply assign some arbitrary chosen numbers to its symbols. We will denote the number assigned to a character c by $\ulcorner c \urcorner$. Thus, e.g. $\ulcorner + \urcorner$ is the unique number assigned to the addition symbol $+$. We call the number $\ulcorner s \urcorner$ the **Gödel code** of the symbol s. We can choose, e.g., the set of even numbers to be the set of codes of variables (then we set $2i$ to be the code of the i-th variable). We then pick up arbitrarily some natural numbers to code the rest of the basic symbols of the syntax of the language of arithmetic.

Note that nothing stops us to speak in PA about codes of symbols from some other language, as long as this language is recursive. E.g. if we want to add three new predicate letters to $\mathscr{L}_{\mathrm{PA}}$ we may simply code them as the first three numbers which are not yet codes of any basic symbols of logic or arithmetic and then speak about finite strings of characters from this enlarged language. Similarly, nothing stops us from adding some infinite families of new symbols as long as these families are recursive. Since we will consider several languages in this thesis, let us assume that from the very beginning we have chosen a coding which embraces all of them. Then all the following remarks about the coding of syntax applies after slight modifications to this extended language although we will state them for $\mathscr{L}_{\mathrm{PA}}$.

We have already defined what codes of symbols are and what sequences are. Then we can define **codes of sequences of characters** simply as sequences of codes of these characters where the latter use of the word "sequence" is understood as above. If s is a sequence of symbols, we will denote its code with $\ulcorner s \urcorner$. Once we know what codes finite strings of symbols are, we may define more elaborate syntactic notions. Let us define for example arithmetical terms *within* PA.

Definition 6. We define the arithmetical formula $\mathrm{Term}_{\mathrm{PA}}(x)$ in the following way: there exists a sequence s of length l such that x is the l-th element of this sequence and for all $i \leq l$ the element s_i is a string of symbols which satisfies one of the following conditions:

1. s_i has length one and is either of the form $\langle \ulcorner 0 \urcorner \rangle$ or $\langle \ulcorner v \urcorner \rangle$ for some variable v.

2. There exists $j < i$ such that s_i is the following string of symbols: $\langle \ulcorner S \urcorner, \ulcorner (\urcorner \frown s_j \frown \langle \ulcorner) \urcorner \rangle$.

3. There exist $j, k < i$ such that s_i has the following form: $\langle \ulcorner (\urcorner \rangle \frown s_j \frown \langle \ulcorner) \urcorner, \ulcorner + \urcorner, \ulcorner (\urcorner \rangle \frown s_k \frown \langle \ulcorner) \urcorner \rangle$.

4. There exist $j, k < i$ such that s_i has the following form: $\langle \ulcorner (\urcorner \rangle \frown s_j \frown \langle \ulcorner) \urcorner, \ulcorner \cdot \urcorner, \ulcorner (\urcorner \rangle \frown s_k \frown \langle \ulcorner) \urcorner \rangle$.

Although the above definition may seem somewhat complicated, it is simply a lengthy way of saying that term is defined inductively as whatever string of characters may be obtained by taking the constant symbol and variables and forming new terms with the symbols $+$, $\cdot$, and S in the usual way.

Following this pattern we may similarly define other inductively defined syntactic concepts. We will now list the syntactic notions which will be used in our thesis.

Definition 7 (Syntactic notions)**.** We define the formulae formalising syntax. Whenever we write that $\phi(x)$ defines the set of objects a, we mean that ϕ formalises the natural recursive definition in analogy to the formula $\mathrm{Term}_{\mathrm{PA}}(x)$ written above.

1. By $\mathrm{Var}(x)$ we mean that x is a code of variable. By convention, this is equivalent to saying that x is even (see our discussion above).

2. $\mathrm{Term}_{\mathrm{PA}}(x)$ defines the set of arithmetical terms.

3. $\mathrm{ClTerm}_{\mathrm{PA}}(x)$ defines the set of closed arithmetical terms, i.e. terms with no free variable.

4. $\mathrm{TermSeq}_{\mathrm{PA}}(x)$ defines the set of sequences of arithmetical terms.

5. $\mathrm{ClTermSeq}_{\mathrm{PA}}(x)$ defines the set of sequences of closed arithmetical terms.

6. $\mathrm{lh}(x, y)$ defines the set of pairs s,l, where s is a sequence and l is its length, i.e. the cardinality of its domain.

7. $\mathrm{Entry}(x, y, x)$ defines the set of triples (s, k, x), where s is a sequence, k is any number no greater than the length of s, and x is the k-th term of the sequence s. In what follows, we will use the functional notation and denote this relation with $(s)_k = x$ or even use $(s)_k$ independently as if it were a term.

8. $\mathrm{FV}(x, y)$ defines the set of pairs (x, a), where x is a term or a formula and a is the set of free variables of x. We will also use the expression $\mathrm{FV}(x)$ as if it were a term to denote the set of free variables of x.

9. $\mathrm{Form}_{\mathrm{PA}}(x)$ defines the set of arithmetical formulae.

10. $\mathrm{Form}^1_{\mathrm{PA}}(x)$ defines the set of arithmetical formulae with exactly one free variable.

11. $\mathrm{Form}^{\leq 1}_{\mathrm{PA}}(x)$ defines the set of arithmetical formulae with at most one free variable.

12. $\mathrm{Sent}_{\mathrm{PA}}(x)$ defines the set of arithmetical sentences.

13. $\mathrm{Val}(x, y)$ defines the set of pairs (x, v), where x is either an arithmetical term or an arithmetical formula and v is a valuation for x, i.e. a function whose domain contains all free variables of x. We will be also using $\mathrm{Val}(x)$ as if it were denoting the class of valuations for x, writing, e.g., $v \in \mathrm{Val}(x)$.

14. $\mathrm{Subst}(x, y, z)$ defines the set of triples (ϕ, t, ψ) such that ϕ is an arithmetical formula with at most one free variable, t is an arithmetical term, and ψ is the formula obtained by substituting the term t for every occurrence of the only variable in ϕ. Following our convention, we will sometimes use the functional notation, i.e., write $\mathrm{Subst}(x, y) = z$ instead of $\mathrm{Subst}(x, y, z)$ or even use $\mathrm{Subst}(x, y)$ as if it were a term.

15. $\mathrm{Substseq}(x, y, z)$ defines the set of triples (ϕ, t, ψ) such that ϕ is an arithmetical formula, t is a sequence of terms of length l, and the formula ψ is a formula obtained by substituting the i-th term in the sequence t for every occurrence of i-th free variable of ϕ for $i \leq l$. We assume that there is a global ordering of all possible variables coming from the ordering of their codes. Additionally, if ϕ has only k variables with $k < l$, then we assume that $\mathrm{Substseq}(\phi, t, \psi) \equiv \mathrm{Substseq}(\phi, t', \psi)$, where $t' = t \restriction \{0, 1, \ldots, k-1\}$, i.e. the sequence t' is t restricted to its first k terms. We will be using $\mathrm{Substseq}(x, y)$ as if it were a term.

16. $\mathrm{val}(x, y)$ defines the set of pairs (t, y), where t is a closed arithmetical term or a sequence of closed arithmetical terms and y is its unique value or the unique sequence of values, respectively. E.g., PA proves that $\mathrm{val}(x, y)$ holds when $x = \ulcorner S(S(S(S(0)))) \urcorner$ and $y = \underline{4}$ or when $x = \ulcorner S(0) + S(0) \urcorner$ and $y = S(0) \times S(S(0))$. Following our convention, we shall use the functional notation, writing $y = x^\circ$. We will also use the expression x° as if it were a term. Officially, this is always an abbreviation which can be eliminated from our formulae.

17. $\mathrm{val}(x, y, z)$ defines the set of triples (t, v, y), where t is an arithmetical term, not necessarily closed, or a sequence of such terms, v is a valuation which comprises all free variables of t and y is the value of t under the valuation v or a sequence of such values. We will be also using expression $v(t)$ to denote this unique element or a sequence of elements y.

18. $\mathrm{Ax}_{\mathrm{PA}}(x)$ defines the set of axioms of PA. More generally, whenever Th is a primitive recursive theory, where a specific primitive recursive axiomatisation is clear from the context, we will write $\mathrm{Ax}_{\mathrm{Th}}(x)$ for the formula which defines the set of axioms of Th.

19. $\mathrm{Prov}_{\mathrm{PA}}(x, y)$ defines the set of pairs (d, ϕ) such that d is the proof in sequent calculus of the sequent $\longrightarrow \phi$ with additional initial sequents allowed of the form $\longrightarrow \eta$, where η is an axiom of PA (we code a sequent arrow $\Gamma \longrightarrow \Delta$ simply as a pair of sequences of formulae). For an arbitrary primitive recursive theory Th, we define $\mathrm{Prov}_{\mathrm{Th}}(x, y)$ in a similar way.

20. If τ is a unary formula, we will write $\mathrm{Prov}_\tau(d, \phi)$ to denote that d is a proof in sequent calculus of the sequent $\longrightarrow \phi$ with additional initial sequents allowed of the form $\longrightarrow \eta$, where every sentence η' obtained by substituting closed terms for eigenvariables in the formula η satisfies the formula τ. In our thesis, τ will be some form of the truth predicate, so $\mathrm{Prov}_\tau(d, \phi)$ typically means that d is a proof of ϕ from true premises.

21. By $\mathrm{Pr}_{\mathrm{PA}}(x)$ we mean $\exists y\, \mathrm{Prov}_{\mathrm{PA}}(y, x)$. We define $\mathrm{Pr}_{\mathrm{Th}}(x)$ for an arbitrary primitive recursive theory Th and $\mathrm{Pr}_\tau(x)$ for a unary formula τ in a similar way.

We will occasionally call the natural numbers corresponding to terms, formulae, etc. **codes of terms, codes of formulae** etc. Most of the time, we will identify syntactic objects with their codes.

The reader may be worried that according to the above definitions given a natural number n, we cannot assign to it a unique "syntactic object" s such that $n = \ulcorner s \urcorner$. For example, one and the same number might be a code for a predicate symbol and a term. Even worse, one and the same number may be a code for a compound arithmetical term and a constant symbol from outside the language of arithmetic. This is indeed the case. The above definitions do not guarantee this kind of uniqueness, but it is also not really needed. Given a number k, it *is* unique what *sequence* this number k is (i.e., what is its length and what are its entries). Therefore, given a number, we can always decode whether it is a string of characters and what characters are they. These characters may be themselves sequences of characters, but we simply do not care about it. According to our definitions, the code of a constant symbol c is never a code of a term which consists only of one constant symbol c, since the latter is the code of $\langle c \rangle$, one-element sequence whose only entry is c.

Let us close this section by quoting the celebrated Gödel theorems. The first of them states that any "reasonably strong" theory is incomplete. Below, we state it in a very special case of Peano Arithmetic.

Theorem 8 (The First Gödel's Theorem)**.** *There exists a sentence $\phi \in \mathscr{L}_{\mathrm{PA}}$ such that*

$$\mathrm{PA} \nvdash \phi \text{ and } \mathrm{PA} \nvdash \neg\phi.$$

Gödel's first theorem is proved by showing that the sentence γ "saying": "γ is not provable in PA" is indeed not provable in PA. Namely, if it were provable, then PA could also prove *that* it is provable in PA. But the latter statement is simply equivalent to $\neg\gamma$, which would yield PA inconsistent. On the other hand, $\neg\gamma$ is also not provable. Namely, it is equivalent to a sentence "γ is provable in PA." So if $\neg\gamma$ were provable, that sentence would be provable as well. But, since $\mathbb{N}$ is a model of PA, the sentence "γ is provable in PA" would hold in $\mathbb{N}$. Then, by "decoding" the object which is claimed in $\mathbb{N}$ to be a proof of γ, we would obtain *an actual* proof of γ in PA. But then there would exist proofs in PA both of γ and $\neg\gamma$ which would yield PA inconsistent.

Of course, in the above sketch, there is a lot of details to be written down and checked in a proper manner. Probably the most mysterious one is to explain what does

it mean that there is a sentence γ which "says of itself" that it satisfies some properties. This is encapsulated in the following lemma.

Lemma 9 (Fixpoint Lemma)**.** *Let Φ be an aribtrary formula in a language $\mathscr{L}$ containing $\mathscr{L}_{\mathrm{PA}}$. Then there exists a sentence α such that*

$$Q \vdash \alpha \equiv \Phi(\ulcorner \alpha \urcorner).$$

Above, we have stated Gödel's First Theorem in a very special case. However, this result admits a number of generalisations. Let us state a version which is reasonably easy to formulate and memorise.

Theorem 10 (Gödel–Rosser Incompleteness Theorem)**.** *Let* Th *be a consistent theory extending* PA *with a primitive recursive axiomatisation. Then there exists a sentence ρ such that*

$$\mathrm{Th} \nvdash \rho \text{ and } \mathrm{Th} \nvdash \neg\rho.$$

We will omit a proof of the above theorem. It builds on the ideas of the proof from Gödel's Theorem but with some nontrivial modifications, since in more general context we cannot resort to the standard model $\mathbb{N}$. A proof of an even stronger statement can be found, e.g., in [Kaye, 1991], Corollary 3.10.

The second Gödel's theorem states that no „reasonably strong" theory proves its own consistency.

Definition 11 (Gödel–Löb provability conditions)**.** Let Th be a theory whose language extends $\mathscr{L}_{\mathrm{PA}}$. We say that a formula $P(x)$ satisfies Gödel–Löb's provability conditions in Th if the following conditions are satisfied:

L 1 For all sentences ϕ, if $\mathrm{Th} \vdash \phi$, then $\mathrm{Th} \vdash P(\underline{\ulcorner \phi \urcorner})$.

L 2 $\mathrm{Th} \vdash P(\underline{\ulcorner \phi \to \psi \urcorner}) \wedge P(\underline{\ulcorner \phi \urcorner}) \to P(\underline{\ulcorner \psi \urcorner})$ for all sentences ϕ, ψ.

L 3 $\mathrm{Th} \vdash P(\underline{\ulcorner \phi \urcorner}) \to P(\underline{\ulcorner P(\underline{\ulcorner \phi \urcorner}) \urcorner})$ for all sentences ϕ.

In the formulation of the above definition, we have tacitly assumed that the language $\mathscr{L}$ of the theory Th may be coded within Th. The canonical example is when $\mathscr{L}$ is a finite expansion of the language $\mathscr{L}_{\mathrm{PA}}$ of Peano Arithmetic. The intended main example of a formula satisfying the Löb's conditions is $\mathrm{Pr}_{\mathrm{PA}}(x)$. It may seem somewhat surprising that the condition L 3 is satisfied. It is indeed not trivial. It actually may fail for certain other "$\mathrm{Pr}_{\mathrm{PA}}$"-like formulae for a number of reasons. Still, we have the following result:

Theorem 12. *The formula $\mathrm{Pr}_{\mathrm{PA}}(x)$ satisfies Gödel–Löb's conditions in* PA.

Now we are ready to state Second Gödel's Theorem. Again, it in fact holds in much greater generality than presented below.

Theorem 13 (The Second Gödel's Theorem)**.** *Let* Th *be any consistent theory extending* Q *with a primitive recursive axiomatisation. Suppose that $P(x)$ satisfies Gödel–Löb's provability conditions. Then*

$$\mathrm{Th} \nvdash \neg P(\underline{\ulcorner 0 \neq 0 \urcorner}).$$

In particular PA *does not prove* $\neg \mathrm{Pr}_{\mathrm{PA}}(\underline{\ulcorner 0 \neq 0 \urcorner})$*, i.e., it does not prove its own consistency.*

Throughout the whole thesis we will make use of a number of conventions to avoid too heavy notation.

Convention 14. We will use the following conventions:

1. We will often drop the "$\ulcorner \urcorner$" symbols and simply identify formulae with their Gödel codes. E.g. we will write $P(\underline{P(\phi)})$ instead of $P(\underline{\ulcorner P(\underline{\ulcorner \phi \urcorner}) \urcorner})$.

2. We will often suppress formulae referring to syntactical operations and write the results of these operations instead. E.g. we will write:

$$\forall x, y \ \Big(\mathrm{Pr}_{\mathrm{PA}}(x \wedge y) \equiv \mathrm{Pr}_{\mathrm{PA}}(x) \wedge \mathrm{Pr}_{\mathrm{PA}}(y)\Big)$$

   rather than

$$\forall x, y, z \ \Big(\mathrm{Conj}(x, y, z) \rightarrow \big(\mathrm{Pr}_{\mathrm{PA}}(z) \equiv \mathrm{Pr}_{\mathrm{PA}}(x) \wedge \mathrm{Pr}_{\mathrm{PA}}(y)\big)\Big),$$

   where $\mathrm{Conj}(x, y, z)$ is a formula defining the triples (x, y, z) such that x, y, z are formulae and z is the conjunction of x and y.

3. For certain formulae $\Phi(x)$ defining sets of syntactical objects we will sometimes write

$$x \in \Phi$$

   instead of $\Phi(x)$. This will be often used to restrict quantification. E.g. we will write

$$\forall t \in \mathrm{Term}_{\mathrm{PA}} \ \mathrm{Pr}_{\mathrm{PA}}(t = t)$$

   rather than

$$\forall t \Big(\mathrm{Term}_{\mathrm{PA}}(t) \rightarrow \mathrm{Pr}_{\mathrm{PA}}(t = t)\Big).$$

   Note that in both formulae above we have already used the conventions which we have listed previously.

1.2. Truth

In this section we shall introduce the formal theories of truth. Further information about them may be found in a monograph [Halbach, 2011]. In particular, one can find there the proofs of all the quoted facts.

1.2.1. Undefinability of truth

The very basic result of truth theory is due to Tarski. It has been formulated already in the seminal work [Tarski, 1995]. Basically, it states that in theories over classical logic, there can be no truth predicate for the whole language. As in the case of Gödel's theorems, we state the result in an admittedly special case.

Theorem 15 (Tarski's undefinability theorem)**.** *Let* Th *be any consistent extension of* Q *over classical logic in a language $\mathscr{L}$. Then there is no formula Θ from the language $\mathscr{L}$ such that for all sentences ϕ in the language $\mathscr{L}$, the following holds:*

$$\mathrm{Th} \vdash \Theta(\phi) \equiv \phi.$$

The equivalences of the above form are called **Tarski's biconditionals**. In this subsection, a formula Θ satisfying all Tarski's biconditionals provably in Th is called **a truth predicate** for the language $\mathcal{L}$.[1] In effect, Tarski's theorem states that there is no theory containing the truth predicate for its own language. Note that in Tarski's theorem we made no assumption that Th is primitive recursive. And in fact we will apply Tarski's theorem to more complicated theories.

The theorem is proved by constructing the liar sentence λ such that

$$\mathrm{Th} \vdash \lambda \equiv \neg T\lambda.$$

That such a sentence exists follows from Fixpoint Lemma 9. Then one readily checks that if T is a truth predicate for the whole language, λ may be shown to be true and not to be true, a contradiction.

Tarski's theorem puts very serious restrictions on the way we can formalise the naïve truth predicate. Let us examine what options are we left with. One option, possible but not viable, is to lift the assumption that our theory contains Q. The assumption may seem somewhat technical. However, this is not really a tenable solution. If we are seriously interested in the enterprise of formally capturing our informal truth predicate for the whole language, then we definitely should assume that our truth theory Th contains Q for at least two reasons. It is intuitively a very modest part of our knowledge of natural numbers, so it would not be philosophically honest to arbitrarily restrict it and, more importantly, in order to make any sense of our discourse about truth of sentences we have to make sense of our talk of sentences in the first place. We must have a reasonable theory of syntax at our disposal and once we assume that we can talk of basic syntactic operations, we can automatically recover an arithmetic theory at least as strong as Q.[2] Therefore, even though we can strictly speaking formulate our truth theory so that it does not extend Q, this would not help in any significant way, since that theory would be still interpretable in our base theory. Therefore this assumption seems to be indispensable in any serious research on truth.

[1] It would be more accurate to say that a formula is a truth predicate *relative to a given theory* but this expression seems to us overly lengthy.

[2] See [Quine, 1946].

Another option we have is to abandon classical logic. We have shown that we cannot define truth predicate for the whole language, since then we could construct the liar sentence, which should be both true and not true. Note however, that the existence of the liar sentence is not a consequence of our truth-theoretic assumptions, but a perfectly general, syntactic phenomenon: we can write down what this sentence is, and prove, using essentially just Q, that it satisfies the equivalence:

$$\lambda \equiv \neg T\lambda.$$

So, if we have any "truth-like" predicate T in our language, we automatically have also a "liar-like" sentence along with it.

Now, intuitively, if we see a sentence λ saying "this sentence is not true", it actually is a quite intuitive option to conclude that λ does not have a well-defined truth value. The existence of the liar sentence and the fact that it can be constructed in a syntactically correct way seems to actually be a good argument against the law of excluded middle. Although embracing non-classical logic seems to us to be a viable option, we will not pursue it further here.

The third option which is left is to construct various approximations to the naïve truth predicate formalising some of its properties and to ask about these approximations those questions that we would like to ask about the truth predicate itself. This third option is somewhat pragmatic in spirit: although we do not have the full naïve truth predicate, we can say much about the concept of truth by obtaining general results showing that some properties of truth predicate must hold, under some very natural assumptions. Basically, this is the way pursued in the study of classical axiomatic theories of truth. We will stick to this ideology for the rest of our work.

1.2.2. Axiomatic theories of truth

Axiomatic theories of truth are intended to "approximate" the non-existent naïve truth predicate Θ satisfying Tarski's biconditionals $\Theta(\phi) \equiv \phi$ for all sentences ϕ. They are typically formed in two steps:

1. We fix some theory B which contains at least a good theory of syntax but typically does not contain any information on semantic notions. We call it the **base theory**.

2. We add to the language of B one new predicate $T(x)$ (with the intended reading "x is a code of a true sentence") and we introduce axioms governing the new predicate. We call it **the truth predicate**.

Then we basically try to understand the properties of theories obtained as the axioms for the predicate T vary. In other words, we try to understand what are the consequences of various assumptions we may have about the behaviour of the truth predicate.

In the investigation on axiomatic theories of truth, one often assumes that the theory B is simply PA. This is mainly for convenience. The only thing which we have to assume about our base theory is that it handles syntax reasonably and PA is much

more than enough to this end, as we have seen in the subsection 1.1.2. Many of the results carry over to theories of truth constructed upon other base theories as well, so the choice of PA as our base theory is in most cases rather harmless. On the other hand, there are situations when the features of the base theory do play some role, so it would be probably best to state abstractly what conditions we require from our base theory and then try to prove results in this abstract setting. Unfortunately, even formulating this abstract theory seems to be a difficult and rather technical task which we will not undertake in this thesis.

We will stick to the terminology introduced above.

Convention 16. Whenever we define some theory which we call **a truth theory** or **a theory of truth** Th, we assume that:

1. The language $\mathscr{L}$ of Th is $\mathscr{L}_{\mathrm{PA}}$ with one new predicate $T(x)$. We denote this language by $\mathscr{L}_{\mathrm{PA}T}$.
2. The theory Th contains PA.

We will define such theories by decribing their axioms governing the truth predicate.

The most basic axiom systems we can assume about our truth predicate are restricted versions of Tarski's biconditionals, where we do not allow to take an arbitrary formula ϕ in the biconditional.

Definition 17. By TB^- we mean a theory of truth, in which the axioms for the truth predicate are biconditionals of the following form:

$$T\phi \equiv \phi,$$

where ϕ is an arithmetical sentence.

The name is of course an abbreviation for "Tarski biconditionals". We may additionally require that the biconditionals hold uniformly.

Definition 18. By UTB^- we mean a theory of truth, in which the axioms for the truth predicate are of the following form:

$$\forall x_1, \ldots, x_n \left(T\phi(\underline{x_1}, \ldots, \underline{x_n}) \equiv \phi(x_1, \ldots, x_n) \right),$$

where $\phi(x)$ is an arithmetical formula.

The name UTB stands for "Uniform Tarski biconditionals".

Another natural and very basic condition one can assume about the truth predicate is that it is compositional. This means for example that a conjunction of two sentences is true if and only if both of the conjuncts are true. Let us formalise this intuition.

Definition 19. By CT^- we mean a theory of truth with the following axioms for the truth predicate:

1. $\forall s, t \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(T(s = t) \equiv s^\circ = t^\circ\Big)$.

2. $\forall \phi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T\neg\phi \equiv \neg T\phi\Big)$.

3. $\forall \phi, \psi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T(\phi \wedge \psi) \equiv T(\phi) \wedge T(\psi)\Big)$.

4. $\forall \phi, \psi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T(\phi \vee \psi) \equiv T(\phi) \vee T(\psi)\Big)$.

5. $\forall v \in \mathrm{Var}\, \forall \phi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}} \Big(T\exists v\phi(v) \equiv \exists x\, T\phi(\underline{x})\Big)$.

6. $\forall v \in \mathrm{Var}\, \forall \phi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}} \Big(T\forall v\phi(v) \equiv \forall x\, T\phi(\underline{x})\Big)$.

The name "CT" stands for "compositional truth". Note that in the above axioms we quantify only over *arithmetical* sentences and formulae. In particular we do not postulate that truth behaves compositionally on sentences which themselves contain the truth predicate. Such a theory would itself be inconsistent, since it would prove all Tarski biconditionals. Note that we understand the word "compositional" so that it implies the condition $T(P(t_1, \ldots, t_n)) \equiv P(t_1{}^\circ, \ldots, t_n{}^\circ)$ for all predicates P and all terms $t_1, \ldots, t_n$ (or even the uniform version of this condition in which terms are quantified).

In fact, much effort has been put in relaxing compositional clauses so that it can consistently be extended to the full language. The most prominent of such attempts is due to Feferman, [Feferman, 1991] building on ideas of [Kripke, 1975]. The theory KF presented there has been extensively investigated in the literature.[3] Its compositional clauses are modelled after strong Kleene's logic rather than classical logic. Namely, the compositional axioms of CT^- implicitly say that the truth predicate respects classical logic. For example, a sentence is not true iff its negation is true. This implies that for any sentence ϕ either this sentence or its negation is true. We can change the compositional clauses, so that they are modelled after some other logics. This has been particularly fruitful in the case of Kleene's logic, as we have mentioned.

Let us now present a theory, whose compositional axioms are that of KF, but with the truth predicate which is *not* self-referential. I.e., the truth predicate in this theory is defined only for arithmetical sentences.

Definition 20. By PT^- we mean a truth theory with the following axioms for the truth predicate:

1. $\forall s, t \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(T(s = t) \equiv s^\circ = t^\circ\Big)$.

2. $\forall s, t \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(T(s \neq t) \equiv s^\circ \neq t^\circ\Big)$.

3. $\forall \phi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T\neg\neg\phi \equiv T\phi\Big)$.

[3]The reader may find the axioms of KF as well as the discussion of this theory in [Halbach, 2011], chapter 15.

4. $\forall \phi, \psi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T(\phi \wedge \psi) \equiv T(\phi) \wedge T(\psi) \Big)$.

5. $\forall \phi, \psi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T\neg(\phi \wedge \psi) \equiv T(\neg\phi) \vee T(\neg\psi) \Big)$.

6. $\forall \phi, \psi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T(\phi \vee \psi) \equiv T(\phi) \vee T(\psi) \Big)$.

7. $\forall \phi, \psi \in \mathrm{Sent}_{\mathrm{PA}} \Big(T\neg(\phi \vee \psi) \equiv T(\neg\phi) \wedge T(\neg\psi) \Big)$.

8. $\forall v \in \mathrm{Var} \forall \phi(v) \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}} \Big(T\exists v \phi(v) \equiv \exists x \, T\phi(\underline{x}) \Big)$.

9. $\forall v \in \mathrm{Var} \forall \phi(v) \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}} \Big(T\neg\exists v \phi(v) \equiv \forall x \, T\neg\phi(\underline{x}) \Big)$.

10. $\forall v \in \mathrm{Var} \forall \phi(v) \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}} \Big(T\forall v \phi(v) \equiv \forall x \, T\phi(\underline{x}) \Big)$.

11. $\forall v \in \mathrm{Var} \forall \phi(v) \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}} \Big(T\neg\forall v \phi(v) \equiv \exists x \, T\neg\phi(\underline{x}) \Big)$.

The name "PT$^-$" stands for "positive truth". Note that in the compositional axioms of PT$^-$ we never say when a sentence is *not true*. We only postulate, when the *negation is true*. This is an extremely important distinction. Intuitively, the former condition is much more restrictive, since it requires us to choose for each pair of sentences $\phi, \neg\phi$ which of them is true, in such a way that the choices are made coherently. The positive axioms do not require us to do anything like this. Of course, this intuition is rather vague, but in the next chapters we will try to show that it is essentially right.

At some point, the reader might have started wondering, why all names of theories which we consider are decorated with the minus sign.

Definition 21. By TB (UTB, CT, PT) we mean a theory of truth obtained via enriching TB$^-$ (UTB$^-$, CT$^-$, PT$^-$, respectively) with the full induction scheme for the formulae containing the truth predicate, i.e. all the axioms of the form:

$$\forall x \Big(\phi(x) \to \phi(Sx) \Big) \longrightarrow \Big(\phi(0) \to \forall x \, \phi(x) \Big),$$

where ϕ is an arbitrary formula, possibly containing the truth predicate and parametres.

In general, a theory Th will always be the same as Th$^-$ with the full induction scheme for the extended language. Also note that theories TB$^-$, UTB$^-$, CT$^-$, PT$^-$ *do have* full induction scheme for the arithmetical formulae as by our convention, they are all extensions of PA.

Many natural theories of truth may be formed via restricting the induction scheme for the formulae of the extended language, e.g. by applying it only to formulae of some particular subclasses. We will introduce some of these theories in further parts of our thesis.

1.3. Models

A good part of our thesis will be devoted to models of axiomatic theories of truth. In this section, we introduce basic notions and facts concerning models of Peano Arithmetic. Obviously, this section cannot replace a proper introduction into model theory of PA, which is a well-established area of research. For such an introduction, the reader is advised to consult, e.g., [Kaye, 1991].

Definition 22. Let $M \models \text{PA}$. We say that M is **nonstandard** if M is not isomorphic to $(\mathbb{N}, 0, S, +, \cdot)$. Any model isomorphic to $\mathbb{N}$ is called **standard**. By a slight abuse of language, we will often call it *the* standard model.

Since writing $(\mathbb{N}, 0, S, +, \cdot)$ is somewhat cumbersome, we will often write simply $\mathbb{N}$ to denote the whole structure together with the standard interpretation of function and constant symbols. We will make so generally when dealing with any model.

Convention 23. Whenever it does not lead to any confusion, we will denote a model and its universe with the same symbol.

Fact 24. *There exists a nonstandard model of* PA.

Proof. Let us consider the following theory Th in $\mathscr{L}_{\text{PA}}$ expanded with a constant c:

$$\text{Th} = \text{Th}(\mathbb{N}) \cup \{c \neq \underline{k} \mid k \in \mathbb{N}\}.$$

Note that above we have already used Convention 23. By compactness, Th has a model $(M, 0^M, c^M, S^M +^M, \cdot^M)$. Since $(M, 0^M, c^M +^M, \cdot^M) \models c \neq \underline{k}$ for any $k \in \omega$, its arithmetical part $(M, 0^M, S^M, +^M, \cdot^M)$ cannot be isomorphic to $\mathbb{N}$. On the other hand, it satisfies PA, since it satisfies even $\text{Th}(\mathbb{N})$. □

The structure of models of PA is in general terribly complicated. Luckily, there are some basic facts which are easy to state and prove. Let us present some of these facts.

Definition 25. Let $I \subset M \models \text{PA}$. We say that I is an **initial segment** of M, if for every $a \in I$, any $b \in M$ such that $M \models b \leq a$ also belongs to I.

Recall that the relation $M \models b \leq a$ to which we have referred above may be defined as follows:

$$M \models \exists x\, a = b + x,$$

so it makes sense to write $a \leq b$ in any model of PA, even though the inequality symbol is not officially in our language.

Proposition 26. *Let M be a model of* Q *(in particular, it can be a model of* PA*). Then M has an initial segment I isomorphic to* $\mathbb{N}$.

Proof. Recall that for any $n \in \mathbb{N}$ there exists a term $\underline{n}$ denoting n in the standard model. Let

$$I = \{\underline{n}^M \mid n \in \mathbb{N}\}.$$

It is enough to check by induction that for all $n \in \mathbb{N}$,

$$Q \vdash \forall x \Big(x \leq \underline{n} \to \big(x = \underline{0} \vee x = \underline{1} \vee \ldots \vee x = \underline{n} \big) \Big).$$

Similarly, one can check by induction on $k \in \mathbb{N}$ that for all $n, k \in \mathbb{N}$

$$Q \vdash \underline{n} + \underline{k} = \underline{n + k} \wedge \underline{n} \times \underline{k} = \underline{n \times k}.$$

This entails that I is isomorphic to $\mathbb{N}$ as an arithmetical structure. □

By slight abuse of language, for every model $M \models \mathrm{PA}$, we shall identify the copy of $\mathbb{N}$ inside M with the actual ω with its natural arithmetical structure. Thus we will say, e.g., "for any $k \in \omega$, $M \models \phi(k)$." This should not lead to any confusion. The fact that we may find a copy of ω inside any model of PA gives rise to the following, simple and important definition:

Definition 27. Let $M \models \mathrm{PA}$ be any model. Let $a \in M$. We say that a is a **standard** element, if a is an element of the unique initial segment of M isomorphic to ω. Otherwise, we say that it is **nonstandard**.

The notion of nonstandard elements will be often coupled with arithmetisation of syntax. Thus, we will speak, e.g. of "nonstadard formulae", "nonstandard terms" and take advantage of the fact that truth predicates allow us to say that these formulae are true or false in a way that enjoys some reasonable properties. Nonstandard formulae exist, since PA proves by induction that for an arbitrary n there is an element x such that $\mathrm{Form}_{\mathrm{PA}}(x)$ holds and $x > n$. In particular, by taking any nonstandard n, we conclude that arbitrarily large nonstandard formulae exist.

Let us recall handy notation which will greatly facilitate speaking about nonstandard syntactic objects. The notation itself is entirely standard.

Definition 28. Let $\phi(x_1, \ldots, x_n)$ be any formula (not necessarily arithmetical). Let M be any model (not necessarily of PA). Then we denote tuples of elements of M satisfying ϕ as follows:

$$\phi(M) = \{(a_1, \ldots, a_n) \in M^n \mid M \models \phi(a_1, \ldots, a_n)\}.$$

Thus e.g., if $M \models \mathrm{PA}$, then by $\mathrm{Form}_{\mathrm{PA}}(M)$ we mean the set of "arithmetical formulae from the point of view of M."

A good part of this thesis is devoted to the study of the interplay between various properties of models of PA and the possible interpretations of truth predicates in these models. Let us introduce the most basic relation in this context.

Definition 29. Let $(M, A_1, \ldots, A_n, f_1, \ldots, f_k)$ be any model (not necessarily in arithmetical signature). Let T be any relation symbol of arity l not in the signature $\mathscr{L}$ of M. Let $T^M \subseteq M$ be any subset of M^l. We call the structure $(M, T, A_1, \ldots, A_n, f_1, \ldots, f_k)$ an **expansion** of M to the signature $\mathscr{L} \cup \{T\}$.

The definition of expansion may be easily adapted so that we allow also expanding with new functions and constants, rather than relations. In our context, the most interesting case is when we expand $M \models \mathrm{PA}$ to a model of some truth theory. To simplify notation, let introduce one more convention.

Convention 30. If $M = (U, A_1, \ldots, A_n)$ is any model with a domain U and

$$(U, P_1, \ldots, P_k, A_1, \ldots, A_n)$$

is its expansion, then we will denote the latter model with $(M, P_1, \ldots, P_k)$. Here A_i, P_i may mean both relation and function symbols.

The above notation will be typically used in our thesis in the following context: if M is a model of PA, then we denote its expansion with a unary predicate T with (M, T). Although this notation introduces some ambiguity, we believe that in practice it proves very handy.

Now, let us introduce several properties of models which will be used to better understand when they are expandable to models of truth theories examined in this thesis. Let us begin with a very basic notion from model theory.

Definition 31. Let M be any model over the language $\mathscr{L}$, let α be any valuation with values in the models M, whose domain are all variables of the language $\mathscr{L}$, and let $p(x, x_1, \ldots, x_n) = \{\phi_i(x, x_1, \ldots, x_n) \mid i \in I\}$ be any set of formulae sharing common variables $x, x_1, \ldots, x_n$. We say that p is a **type** (with parametres) over the model M under the valuation α, if for all finite subsets $I_0 \subseteq I$

$$M \models \exists x \bigwedge_{i \in I_0} \phi_i(x, x_1, \ldots, x_n)[\alpha].$$

Usually, instead of considering valuations and taking advantage of the fact that formulae ϕ_i may share free variables, we simply say that the formulae in p are allowed to have some parametres. It is convenient to view them as additional constants denoting elements of M. In what follows, we will completely omit mentioning the valuation in the definition of type and speak as if the formulae in a given type simply contained new constants. Note however, that in the context of strong enough truth theories we do not have to introduce some separate codes for the formulae in the language expanded with constants, since we may represent these formulae with codes of arithmetical formulae containing nonstandard terms. I. e., whenever we see a formula $\phi(x, c_1, \ldots, c_n)$ where $c_1, \ldots, c_n$ are constants denoting elements $a_1, \ldots, a_n \in M$ then we may treat the code $\ulcorner \phi(x, \underline{a_1}, \ldots, \underline{a_n}) \urcorner$ (which, in M is a nonstandard formula in the purely arithmetical language) as the code of $\phi(x, c_1, \ldots, c_n)$.

So, a type is some set of formulae possibly with new constants, denoting elements of M, whose all finite subsets are realised by some elements (possibly different for different subsets).

Definition 32. Let M be any model. Let $p(x)$ be any type over M under a valuation α. We say that p is **realised** in M if there exists $a \in M$ such that for all $\phi \in M$,

$$M \models \phi(a),$$

i.e., there exists a valuation α' differing from α at most at the variable x such that $\alpha'(x) = a$, and

$$M \models \phi[\alpha'].$$

If p is not realised in M, we say that it is **omitted**.

So a type p is realised simply if all formulae $\phi \in p$ are satisfied jointly by the same element. Let us present an example of what does it mean to omit a type.

Example 33. Let $p = \{x > \underline{n} : n \in \omega\}$. Then p is a type without parametres over any model of PA which is omitted in $\mathbb{N}$ and realised in every other model.

The types may be terribly wild as sets of formulae. A typical example of such a type is the set of formulae which are satisfied in a given model by some fixed element. On the other hand, the type introduced in the above example is very simple and has fairly regular structure. We would like to capture this distinction between more or less complicated types. This lead to the following definition:

Definition 34. Let p be a type over M in a language $\mathscr{L}$. Suppose that p consists of countably many formulae. Then we say that p is **recursive**, if the set of codes of formulae in M is recursive.

Note that we have implicitly assumed in the above definition that the language $\mathscr{L}$ is given together with some fixed coding.

Definition 35. Let M be any model. We say that M is recursively saturated if any recursive type p over M is realised in M.

The notion we have just defined is not vacuous. It is relatively easy to see, via a routine elementary chain argument that recursively saturated models of PA indeed exist. Actually, a much stronger fact holds whose formulation may be found, e.g., in [Kaye, 1991], Proposition 11.4. Below we give its special version which will suffice for our purposes.

Fact 36. *For any model $M \models$ PA, there exists a recursively saturated model $N \models$ PA such that $M \preceq N$.*

We have already seen a model of PA which is not recursively saturated. Namely, the standard model. Besides that, there are lots of nonstandard models which are not recursively saturated. One can think that recursive saturation implies that a given model is "rich" or "abundant" in elements. Let us introduce the class of particularly "thin" models.

Definition 37. Let M be any model of PA. By $K(M)$ we mean the substructure of M whose domain is formed by taking all first-order definable elements of M, i.e. the set of all $a \in M$ such that there exists $\phi \in \mathrm{Form}_{\mathrm{PA}}^{\leq 1}$ with

$$M \models \phi(a) \wedge \forall x \Big(\phi(x) \rightarrow x = a \Big).$$

We call models of the form $K(M)$ for some M **prime models**.

Suppose that $A \subseteq M$. By $K(M, A)$ we mean the substructure of M, whose domain is the set of elements of M definable by first-order formulae with parametres from A. If A happens to be a finite set $\{a_1, \ldots, a_n\}$ we write $K(M, a_1, \ldots, a_n)$ meaning $K(M, \{a_1, \ldots, a_n\})$.

From what we have written above, it is not clear that the theory of the model $K(M, A)$ depends in general on the theory of M in some interesting way. It is not even clear that $K(M, A)$ is a model of PA. Fortunately, the following fact holds:

Fact 38. *Let $M \models$ PA and let A be a subset of M, possibly empty. Then*

$$K(M, A) \preceq M,$$

i.e., $K(M, A)$ is an elementary substructure of M.

The proof of the above fact may be found in [Kaye, 1991], Theorem 8.1. We will show that prime models are indeed "thin."

Proposition 39. *Suppose that $K = K(M, a_1, \ldots, a_n)$ for some model $M \models$ PA. Then K is not recursively saturated.*

Proof. Fix any model $K = K(M, a_1, \ldots, a_n)$, where $M \models$ PA. Let us consider the following type over K:

$$p(x) = \{ \ulcorner \phi(\underline{a_1}, \ldots, \underline{a_n}) \urcorner \in x \equiv \phi(a_1, \ldots, a_n) \mid \ulcorner \phi \urcorner \in \mathrm{Form}_{\mathrm{PA}}^{\leq n} \}.$$

It states that the element x codes the set of all sentences with parametres $a_1, \ldots, a_n$ which are true in K. We will show that $p(x)$ is omitted in K.

If $p(x)$ were not omitted, then it would be realised by some element $a \in K$. Since, $K = K(M, a_1, \ldots, a_n)$, the element a is definable in M with parametres. I.e., there exists a formula ψ such that

$$M \models \psi(a, a_1, \ldots, a_n) \wedge \forall x \Big(\psi(x, a_1, \ldots, a_n) \rightarrow x = a \Big).$$

Note that by elementarity, the same holds also in K. Now consider the expansion $(K, a'_1, \ldots, a'_n)$ of K by constants $a'_1, \ldots, a'_n$ interpreted as elements $a_1, \ldots, a_n$. Since a realises p in K, for any sentence ϕ in the language $\mathscr{L}_{\mathrm{PA}} \cup \{a'_1, \ldots, a'_n\}$ the following holds:

$$\mathrm{Th}(K, a_1, \ldots, a_n) \models \phi \equiv \exists x \Big(\psi(x, a'_1, \ldots, a'_n) \wedge \ulcorner \phi \urcorner \in x \Big).$$

This contradicts Tarski's Theorem 15 for the theory $\mathrm{Th}(K, a_1, \ldots, a_n)$. □

There is a weakening of the notion of recursive saturation which turns out to be very natural and which is quite useful in the context of truth theories.

Definition 40. Let $M \models \mathrm{PA}$. We say that a model M is **short recursively saturated**, if M realises every recursive type $p(x)$ containing the formula $x < a$ for some parameter $a \in M$.

So the model is short recursive saturated if it realises every type, which is finitely satisfied *below a given element* a. Moreover, we require that the type is also satisfied by an element smaller than a. Now, the proof of Proposition 39 may be slightly modified by requiring that the type p contains the formula $x < b$ for some additional parameter $b \in M$. This gives us a related result.

Proposition 41. *Let $K = K(M, a_1, \ldots, a_n)$ for some $M \models \mathrm{PA}$. Then K is not short recursively saturated.*

On can wonder whether there are short recursively saturated models which are not recursively saturated. This is indeed the case. It can be easily checked that $\mathbb{N}$ actually happens to be short recursively saturated. In fact, there exist also nonstandard models of PA which are short recursively saturated but not recursively saturated. The proof of this fact is not difficult, but it requires some more advanced facts about models of PA. Since, we will not use these facts, we shall omit the proof.

Proposition 42. *There exists a nonstandard model $M \models \mathrm{PA}$ such that M is short recursively saturated but not recursively saturated.*

1.4. Tools

In this section we will describe some lemmata which will recur in our thesis. In our research, it turned out that a couple of techniques may be repeatedly applied to several distinct problems concerning strength of truth theories. We tried to isolate the lemmata which we found particularly useful. We believe that they may find some applications in further research on axiomatic truth theories.

1.4.1. Generalised commutativity

The lemma to be presented in the current subsection is close to trivial. Suppose that (M, T) is any model of CT^- and let $\psi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}(M)$. Consider the following (quite random) formula $T\alpha$:

$$T\Big((\psi(\underline{a}) \wedge \neg\psi(\underline{b})) \vee \underline{0} = S(\underline{c})\Big).$$

One may check that by a couple of applications of compositional axioms of CT^- the sentence $T\alpha$ is equivalent to:

$$(T\psi(\underline{a}) \wedge \neg T\psi(\underline{b})) \vee \underline{0} = S(\underline{c}).$$

In other words, since the truth predicate commutes with (or, more accurately, is distributive over) any quantifier and connective, it commutes also with whole finite syntactic trees. The only thing which we have to do is to state this in a proper way. The rest of this section is devoted to this aim.

Definition 43. Let ϕ be any formula in a language containing a unary predicate $P(x) \notin \mathscr{L}_{\mathrm{PA}T}$ and let Θ be any formula of the same language with precisely one free variable. Then by

$$\phi[\Theta]$$

we mean a formula obtained via substituting in ϕ for any term t (not necessarily closed), the formula $\Theta(t)$ for each occurrence of $P(t)$, possibly renaming bounded variables in Θ, so as to avoid clashes.

We will denote the language extending $\mathscr{L}_{\mathrm{PA}}$ with a fresh unary predicate $P(v)$ by $\mathscr{L}_{\mathrm{PA}P}$.

Let us illustrate the above definition with an example, which should make it completely clear.

Example 44. Let ϕ be equal to:

$$P(x+y) \wedge x > 0 \wedge \exists x \, \forall v \; P(v).$$

Let $\Theta(w) = \exists v \; v = w$. Then $\phi[\Theta]$ equals to:

$$\exists v \; (v = x + y) \wedge x > 0 \wedge \exists x \, \forall v \exists z (z = v).$$

In our applications, Θ will be usually some formula behaving similarly to the truth predicate. The presence of terms under the P predicate is then quite a nuisance, so we will explicitly postulate that this does not happen.

Definition 45. Let ϕ be any formula containing the unary predicate $P(x)$. We say that ϕ is **semirelational** if the predicate P is applied in ϕ only to variables.

The following elementary lemma allows us to not worry about terms under the P predicate.

Lemma 46. *Let ϕ be any formula. Then ϕ is equivalent in classical logic to a semirelational formula.*

To see that the above lemma holds, note that we can simply replace in ϕ any subformula $P(t)$ with $\exists v \; v = t \wedge P(v)$ for some variable v. Let us define one more notion. It is not at all deep, but it allows us to speak in a more convenient manner about truth predicates which are compositional only to some degree.

Definition 47. Let M be an arbitrary model of PA, $\phi \in \mathrm{Form}_{\mathrm{PA}}(M)$ and let Θ be a unary predicate with a fixed interpretation in M. We say that Θ is **fully compositional at** ϕ in a model (M, Θ^M) if and only if the following conditions are satisfied:

1. If there are $s, t \in \mathrm{ClTerm}_{\mathrm{PA}}(M)$ such that $M \models \phi = (s = t)$, then $(M, \Theta^M) \models \Theta(\phi) \equiv s^\circ = t^\circ$.

2. If there is $\psi \in \mathrm{Sent}_{\mathrm{PA}}(M)$ such that $M \models \phi = \neg\psi$, then $(M, \Theta^M) \models \Theta(\phi) \equiv \neg\Theta(\psi)$.

3. If there exist $\psi, \xi \in \mathrm{Sent}_{\mathrm{PA}}(M)$ such that $M \models \phi = \psi \wedge \xi$, then $(M, \Theta^M) \models \Theta(\phi) \equiv \Theta(\psi) \wedge \Theta(\xi)$.

4. If there exist $\psi, \xi \in \mathrm{Sent}_{\mathrm{PA}}(M)$ such that $M \models \phi = \psi \vee \xi$, then $(M, \Theta^M) \models \Theta(\phi) \equiv \Theta(\psi) \vee \Theta(\xi)$.

5. If there exists $v \in \mathrm{Var}(M)$ and $\psi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}(M)$ such that $M \models \phi = \exists v \psi(v)$, then $(M, \Theta) \models \Theta(\phi) \equiv \exists x \Theta\psi(\underline{x})$.

6. If there exists $v \in \mathrm{Var}(M)$ and $\psi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}(M)$ such that $M \models \phi = \forall v \psi(v)$, then $(M, \Theta) \models \Theta(\phi) \equiv \forall x \Theta\psi(\underline{x})$.

Definition 48. Suppose that M is a model of PA and Θ is a unary predicate with a fixed interpretation in M.

- If $\Gamma \subset M$, we say that Θ is compositional at Γ if it is compositional at all $\phi \in \Gamma \cap \mathrm{Form}_{\mathrm{PA}}(M)$.

- Let $\psi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}(M)$. Let $\phi \in \mathrm{Form}_{\mathrm{PA}P}(M)$ be an arbitrary formula. We say that Θ is compositional **across** (ϕ, ψ) if Θ is compositional at every sentence resulting from substituting numerals in the formula $\xi[\psi]$, where $\xi \in \mathrm{Form}_{\mathrm{PA}P}$ is a subformula of ϕ different from $P(v)$.

Let us comment on the second item of the definition, since it is admittedly somewhat technical. However, the intuition behind this definition is really simple. Several times, we will see the following situation: there is a (nonstandard) formula ψ and a standard formula $\phi \in \mathscr{L}_{\mathrm{PA}P}$. We then want to use the fact that a truth-like predicate Θ behaves like a compositional predicate at every subformula of $\phi[\psi]$, until we hit ψ. We treat ψ as a primitive predicate and we do not really care whether we can push the predicate Θ further down the syntactic tree of ψ using compositional rules. The definition above exactly spells out that this situation happens. For example, literally any predicate Θ is compositional across (ϕ, ψ) if ϕ is of the form $P(v)$. Note that given two formulae η, ψ, the presentation of η as $\phi[\psi]$ is in general not unique if it exists. Namely, an occurrence of the formula ψ in η may be induced both by substituting it for a variable P in ϕ and by the fact that a copy of ψ was already present within the formula ϕ. Therefore, in our definition we explicitly mention both ψ and ϕ in which ψ is to be inserted. Now let us introduce on more piece of very useful notation.

Definition 49. Let M be any model of PA. Let $\psi \in \mathrm{Form}_{\mathrm{PA}}(M)$. Let Θ be an arbitrary unary formula (possibly not arithmetical). Then

$$\Theta * \psi(x_1, \ldots, x_n) = \Theta(\psi(\underline{x_1}, \ldots, \underline{x_n})).$$

The notation introduced above will be used for Θ's resembling truth predicate. After these preparatory steps, we are finally ready to state our lemma.

Lemma 50 (Generalised commutativity lemma)**.** *Let M be any model of* PA*, $\psi \in \mathrm{Form}^1_{\mathrm{PA}}(M)$. Let $\phi(x_1 \ldots, x_n)$ be an arbitrary standard semirelational formula of language $\mathscr{L}_{\mathrm{PA}P}$ and let Θ be an arbitrary unary predicate. Suppose that Θ is compositional across (ϕ, ψ) in (M, Θ). Then*

$$(M, \Theta) \models \forall x_1, \ldots, x_n \Big(\Theta(\phi[\psi](\underline{x_1}, \ldots, \underline{x_n})) \equiv \phi[\Theta * \psi](x_1, \ldots, x_n)\Big).$$

Note that in the above lemma ϕ is an actual, standard formula. So it really says that compositionality allows us to distribute the truth predicate over finite syntactic trees. This is also basically the proof of the lemma. Before we proceed to the proof, let us state one corollary. From Lemma 50, we conclude that the compositional axioms for the truth predicate allow us to recover Tarski's biconditionals.

Proposition 51. *The theory* UTB$^-$ *of uniform Tarski's biconditionals is contained in* CT$^-$*. Moreover, it is contained in* PT$^-$*.*

Proof. The first part follows by Lemma 50 for $\Theta = T$ by considering an arbitrary arithmetical formula ϕ as a formula of extended language with no occurrence of the new predicate $P(v)$. In particular, any such ϕ is semirelational.

The "moreover" part follows, since one can simultaneously check by induction on complexity of formulae that in PT$^-$ the truth predicate $T(x)$ is compositional at standard arithmetical formulae and all standard arithmetical formulae are total, i.e., for any standard $\phi(v_1, \ldots, v_n) \in \mathrm{Form}_{\mathrm{PA}}$ the following holds:

$$\forall t_1, \ldots, t_n \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(T\phi(\underline{t_1}, \ldots, \underline{t_n}) \vee T\neg\phi(\underline{t_1}, \ldots, \underline{t_n})\Big).$$

□

Now, we can prove Lemma 50. As noted before, we basically have to spell out in some detail that compositionality allows us to push the predicate Θ finitely many levels down the syntactic tree of any formula ϕ.

Proof. We prove the lemma by induction on complexity of ϕ. If ϕ is atomic, then since it is semirelational, our ϕ equals to $s = t$ for some arithmetical terms s, t or $\phi = P(x)$ for some variable x.

If $\phi = \big(s(x_1, \ldots, x_n) = t(x_1, \ldots, x_n)\big)$, then $\phi[\psi] = \phi$ and

$$\begin{aligned} \Theta(\phi(\underline{x_1}, \ldots, \underline{x_n})) &= \Theta(s(\underline{x_1}, \ldots, \underline{x_n}) = t(\underline{x_1}, \ldots, \underline{x_n})) \\ &\equiv s(\underline{x_1}, \ldots, \underline{x_n})^\circ = t(\underline{x_1}, \ldots, \underline{x_n})^\circ \\ &\equiv s(x_1, \ldots, x_n) = t(x_1, \ldots, x_n). \end{aligned}$$

The first equivalence follows from compositionality of Θ. Note that ϕ is standard, so the last expression makes sense.

If $\phi = P(v)$, then $\phi[\psi] = \psi(v)$. Analogously, $\phi[\Theta * \psi] = \Theta * \psi$. But by definition

$$\Theta\psi(\underline{x}) = \Theta * \psi(x).$$

This proves our lemma for the atomic sentences.

The induction step for propositional connectives is straightforward. So let us suppose that the lemma holds for $\eta(x_1, \ldots, x_n, v)$ and let us prove it for $\phi = \exists v\eta$. The following equivalences hold:

$$\begin{aligned}
\Theta(\phi[\psi](\underline{x_1}, \ldots, \underline{x_n})) &= \Theta(\exists v\eta[\psi](\underline{x_1}, \ldots, \underline{x_n}, v)) \\
&\equiv \exists v\Theta(\eta[\psi](\underline{x_1}, \ldots, \underline{x_n}, \underline{v})) \\
&\equiv \exists v\eta[\Theta * \psi](x_1, \ldots, x_n, v) \\
&= \phi[\Theta * \psi](x_1, \ldots, x_n).
\end{aligned}$$

The case of the universal quantifier is fully analogous. □

1.4.2. Recovering induction from internal induction

Let us begin this section with a basic result on CT which might help better understand a simple but somewhat technical lemma which we discuss in this part.

Proposition 52. CT *proves the following principle of the* ***axiomatic soundness*** *of* PA*:*

$$\forall\psi\Big(\mathrm{Ax}_{\mathrm{PA}}(\psi) \to T(\phi)\Big). \tag{AS}$$

Proof. Axioms of PA comprise the infinite scheme of induction and finitely many other axioms. Using Proposition 51, we can show that these finitely many axioms are true. So it suffices to show that for all ϕ, if ϕ is an instance of the induction scheme, then it is true. Note that for any model $(M, T) \models$ CT and any $\psi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}(M)$, the following holds in (M, T), since it is an instance of the induction scheme for $\mathscr{L}_{\mathrm{PA}T}$:

$$\forall x\ \Big(T * \psi(x) \to T * \psi(Sx)\Big) \longrightarrow \Big(T * \psi(0) \to \forall x\ T * \psi(x)\Big).$$

Then by a few applications of the compositional axioms, we conclude that

$$T\Bigg(\forall x\ \Big(\psi(x) \to \psi(Sx)\Big) \longrightarrow \Big(\psi(0) \to \forall x\ \psi(x)\Big)\Bigg).$$

□

Let us mention that the formulae of the form

$$\forall x\ \Big(T * \psi(x) \to T * \psi(Sx)\Big) \longrightarrow \Big(T * \psi(0) \to \forall x\ T * \psi(x)\Big)$$

are instances of the **internal induction scheme**. We will return to it in further chapters.

Generalised commutativity may be used to prove the following result which is close to trivial but still surprisingly fruitful.

Lemma 53 (Internal–External Lemma)**.** *Let* (M, Θ) *be an arbitrary model with* $M \models \mathrm{PA}$*. Let* $\psi \in \mathrm{Form}_{\mathrm{PA}}^{\leq 1}$ *and let* $\phi \in \mathscr{L}_{\mathrm{PA}P}$ *be a standard unary formula. Suppose that:*

- $(M, \Theta) \models \forall x \left(\Theta * \phi[\psi](x) \to \Theta * \phi[\psi](Sx)\right) \longrightarrow \left(\Theta * \phi[\psi](0) \to \forall x\, \Theta * \phi[\psi](x)\right)$.
- Θ *is compositional across* (ϕ, ψ).

Then

$$(M, \Theta) \models \forall x \left(\phi[\Theta * \psi](x) \to \phi[\Theta * \psi](Sx)\right) \longrightarrow \left(\phi[\Theta * \psi](0) \to \forall x\, \phi[\Theta * \psi](x)\right).$$

Let us comment upon the above lemma before we proceed to proof. We will apply the lemma in circumstances when Θ is thought of as some form of a truth predicate. Then the above lemma essentially allows us to derive the full induction scheme for the predicate $\Theta * \psi$ from the full *internal* induction for the predicate Θ. The lemma holds also if we allow parametres in the formula ϕ with an obvious adaptation of the proof.

Proof. Suppose that in (M, Θ) the following formula holds:

$$\forall x \left(\Theta * \phi[\psi](x) \to \Theta * \phi[\psi](Sx)\right) \longrightarrow \left(\Theta * \phi[\psi](0) \to \forall x\, \Theta * \phi[\psi](x)\right)$$

By Generalised Commutativity Lemma 50, this is equivalent to:

$$(M, \Theta) \models \forall x \left(\phi[\Theta * \psi](x) \to \phi[\Theta * \psi](Sx)\right) \longrightarrow \left(\phi[\Theta * \psi](0) \to \forall x\, \phi[\Theta * \psi](x)\right).$$

This concludes our proof. □

As we have written at beginning of this section, the Lemma will be mostly used to obtain induction for the truth-like predicates *from* the fact that according to this truth predicate, some instances of the induction scheme are true. But notice that this actually works both ways.

Chapter 2

Proof-theoretic strength

In this chapter, we shall investigate the proof-theoretic strength of various compositional theories of truth. Most importantly, we will try to understand, which natural compositional theories of truth are non-conservative over PA, i.e., prove more arithmetical consequences than PA itself.

2.1. Syntactic conservativity of truth theories

In our thesis, we deal with the following general problem: Which natural properties of the truth predicate make the truth theory stronger than its base theory? Let us introduce one of the basic explications of what one can mean by "stronger."

Definition 54. Let $\mathrm{Th} \subseteq \mathrm{Th}'$ be any two theories and let $\mathscr{L}_{\mathrm{Th}}$ be the language of the first theory. We say that Th' is **syntactically (proof-theoretically) conservative** over Th if for any sentence $\phi \in \mathscr{L}_{\mathrm{Th}}$

$$\mathrm{Th}' \vdash \phi \text{ if and only if } \mathrm{Th} \vdash \phi.$$

Obviously, the above notion is interesting only if the language of Th' strictly extends the language of Th. So, intuitively, conservativity means that although Th' introduces new concepts, these concepts do not imply any new substantial consequences. As long as we are interested only in the consequences expressible in the language $\mathscr{L}_{\mathrm{Th}}$, these new notions are fully dispensable. In this chapter, we will simply speak of **conservativity**, meaning syntactic conservativity, since it is not until the next chapter that we introduce other variants of this notion.

The notion of conservativity easily generalises from a tool to divide theories into weak and strong into a measure to compare strength of theories.

Definition 55. Let $\mathrm{Th}, \mathrm{Th}'$ be any two theories whose languages contain some fixed language $\mathscr{L}$. We say that Th' is **syntactically at least as strong as** Th over $\mathscr{L}$ if for any sentence ϕ from the language $\mathscr{L}$, if $\mathrm{Th} \vdash \phi$, then $\mathrm{Th}' \vdash \phi$. We denote this relation with $\mathrm{Th}' \geq_{\mathscr{L}} \mathrm{Th}$.

This notion of strength allows us to define also the relations of begin *strictly* stronger than or being of the same strength in the obvious way. Since the phrase "Th is syntactically as strong as Th$'$ over $\mathscr{L}$" is somewhat lengthy we will skip the word "syntactically" until we discuss other notions of strength. We will also skip the mention of the theory $\mathscr{L}$, always assuming that $\mathscr{L}$ from the above definition is the language of our base theory B for the theory of truth. Since B is assumed to be PA, we will sometimes say that Th$'$ is **arithmetically stronger** than Th to denote the relation introduced above. In this chapter, we will be writing simply Th$' \geq$ Th, rather than Th$' \geq_{\mathscr{L}_{\mathrm{PA}}}$ Th.

Generally, truth theories are significantly stronger than their base theories. Let us present the canonical example along these lines which was already anticipated by Tarski.[1]

Theorem 56. CT *is not conservative over* PA.

Proof. Working in CT, we check that for any proof d in sequent calculus with initial sequents of the form

$$\longrightarrow \phi,$$

where ϕ is an axiom of PA, for any sequent $\Gamma \longrightarrow \Delta$ in d, and any substitution of closed terms for eigenvariables in that sequent, if all formulae from Γ are true under this substitution, then some formula in Δ is true under this substitution.[2] We prove the claim by induction on size of the derivation d.

If the proof d consist only of an initial sequent $\longrightarrow \phi$ with $\phi \in \mathrm{Ax}_{\mathrm{PA}}$, then the claim follows by Proposition 52. If d consist only of one sequent of the form $\phi \longrightarrow \phi$, then the claim follows by pure first-order logic. If the proof results by applying one of the rules of the sequent calculus to a subderivation d_1 or subderivations d_1, d_2, then the claim readily follows by induction hypothesis and the compositional axioms of CT$^-$.

In case of the quantifier axioms we have to check that we can infer $T\exists v\phi(v)$ from any sentence of the form $T\phi(t)$, where t is a closed term and that we can infer any sentence of the form $T\phi(t)$ from the sentence $T\forall v\ \phi(v)$. Both of these facts follow from the compositional axioms for the quantifiers, the fact that provably in PA every closed term has a value equal to the value of some numeral, and the fact that, provably in CT, if two sentences differ only by substitution of closed terms with equal values, then they are either both true or both false.

Now, suppose that there exists a proof d in sequent calculus from the axioms of PA, whose conclusion is

$$\longrightarrow \neg 0 = 0.$$

Then by the above claim, we would have $T(\neg 0 = 0)$. However, this is impossible by the compositional axioms of CT. Thus there is no such proof d.

[1] The variant of the proof we present here can be found, for instance, in [Halbach, 2011], Theorem 8.39 and Corollary 8.40 although the reasoning itself is well-known and it seems that the theorem itself was anticipated in [Tarski, 1995], *Postscript*.

[2] Note, that this is actually a slight abuse of language. It makes no sense to say that a *formula* is true under a substitution. What we actually mean, is that a sentence resulting from substituting in that formula numerals corresponding to the values of free variables under the given valuation is true.

In effect, we have shown that there is no proof d of the sentence $\neg 0 = 0$ from the axioms of PA, i.e., we have shown Con(PA). This is an arithmetical sentence which is unprovable in PA itself by Gödel's Second Theorem (Theorem 13). □

Since one of the observations used in the above proof is not quite obvious, let us isolate it as a separate lemma.

Lemma 57 (Extensionality Lemma)**.** CT *proves the following* ***extensionality principle****:*

$$\forall \phi \in \mathrm{Form}_{\mathrm{PA}}^{\leq 1} \forall s, t \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(s^\circ = t^\circ \to T\phi(s) \equiv T\phi(s) \Big). \tag{EXT}$$

Proof. We check by a straightforward induction on complexity of formulae that for any $\phi \in \mathrm{Form}_{\mathrm{PA}}$ and any substitution of terms for free variables of ϕ, the truth of the resulting sentence depends only on the values of the substituted terms rather than the terms themselves. □

One can check that in the both above proofs we have actually used only Π_1-induction for the compositional truth predicate. One has to check that axiomatic soundness may be obtained with this restricted amount of induction, but we will soon show in Section 2.2 that actually even Δ_0-induction suffices to this end. Let CT_1 be CT^- with the induction scheme for Π_1 formulae containing the arithmetical truth predicate. Then we have the following corollary:

Corollary 58. CT_1 *is not conservative over* PA.

Corollary 59. CT_1 *proves the extensionality principle EXT.*

Another classical result on conservativity gives an example of a natural *weak* theory of truth.[3]

Theorem 60. UTB *(and hence* TB*) is conservative over* PA.

Proof. Take any proof d in UTB and let $\phi_1 \dots, \phi_n$ be the instances of the uniform disqutation scheme which occur in d. Every ϕ_i is of the shape

$$\forall x_1, \dots, x_k \Big(T\psi_i(\underline{x_1}, \dots, \underline{x_k}) \equiv \psi_i(x_1, \dots, x_k) \Big)$$

for some ψ_i. Let N be the minimal number such that $\psi_j \in \Sigma_N$ for all $j \leq n$. Then let d' be d with all instances of T predicate replaced with the the arithmetical truth predicate Tr_N. Let d^* be d' with every instance of the uniform disquotation scheme for Tr_N preceded with its proof in PA. That such proofs exist follows from our choice of N. We readily check that d^* is a valid proof, i.e., that all axioms containing the predicate T which occur in d are satisfied with T replaced by Tr_N. For the disquotation axioms, this is clear by construction. For the induction axioms, this follows by the fact that Tr_N is an arithmetical formula. □

[3]The result is already implied by the proof of Theorem III in [Tarski, 1995].

By inspection of the above proof, we may conclude that UTB has at most polynomial speed-up over PA. I.e., there exists a polynomial p such that for every proof d of an arithmetical sentence in PA there exists a proof d' in UTB with the same conclusion such that

$$d' \leq p(|d|),$$

where $|d|$ is the number of symbols in d. An easy argument for at most polynomial speed-up follows by the observation that rather than the predicate Tr_N we could have used a very simple formula $\Theta(x)$ defined as: "either there exist terms $t_1, \ldots, t_{k_1}$ such that x is obtained by substituting $t_1, \ldots, t_{k_1}$ to $\psi_1(v_1, \ldots, v_{k_1})$ and $\psi_1(t_1{}^\circ, \ldots, t_{k_1}{}^\circ)$ or there exist terms $t_1, \ldots, t_{k_2}$ such that x is obtained by substituting $t_1, \ldots, t_{k_2}$ to $\psi_2(v_1, \ldots, v_{k_2})$ and $\psi_2(t_1{}^\circ, \ldots, t_{k_2}{}^\circ)$ or ... or there exist terms $t_1, \ldots, t_{k_n}$ such that x is obtained by substituting $t_1, \ldots, t_{k_n}$ to $\psi_n(v_1, \ldots, v_{k_n})$ and $\psi_n(t_1{}^\circ, \ldots, t_{k_n}{}^\circ)$." Then the formula Θ is polynomial in the size of $\phi_1, \ldots, \phi_n$ and the proof that Θ satisfies uniform disquotation for $\phi_1, \ldots, \phi_n$ is also polynomial in the size of these formulae.

The proof presented above leaves the impression that disquotational truth theories are somewhat trivial. This impression is not quite correct, as one can recover surprisingly much strength from purely disquotational axioms in the setting where truth predicate is self-referential.[4] However, it is at least hard to see a similarly simple proof of conservativity of compositional truth theories, even if we drop the assumption that the truth predicate satisfies the full induction scheme. Therefore, one could actually wonder whether the compositional truth theory CT^- with completely no induction for the extended language is conservative over PA. It turns out that the answer is positive, although nontrivial.

Theorem 61 (Kotlarski–Krajewski–Lachlan, Enayat–Visser, Leigh)**.** CT^- *is conservative over* PA.[5]

The methods employed to show that CT^- is conservative may be used to obtain a much stronger result which is somewhat surprising. As we have seen in the proof of Theorem 56, there are two ingredients to the proof of nonconservativity of CT. First,

[4] See, [Halbach, 2009], Theorem 5.1 for the example of PUTB—a disquotational truth theory as strong as KF. See also [Cieśliński, 2011] where non-uniform version of the above theory is showed to be weak.

[5] Kotlarski, Krajewski, and Lachlan have shown a related result for something resembling satisfaction predicate, rather than truth predicate in [Kotlarski et al., 1981]. It is, however, not quite obvious, how their proof generalises to the latter setting. Then a neat model-theoretic argument for the conservativity of truth theory over PA formulated in the *relational* language (i.e. a language with no function symbols) has been presented by Enayat and Visser, see [Enayat and Visser, 2015], Theorem 3.2. Conservativity of CT^- over PA as defined in this thesis has been shown independently by Leigh in [Leigh, 2015], Theorem 1, using proof-theoretic methods.

The history of Theorem 62 is similar. All three groups of authors have found much more general result of which conservativity of $\mathrm{CT}^- + \mathrm{AS}$ is a corollary. This has been formulated by Kotlarski, Krajewski, and Lachlan for something resembling satisfaction rather than truth predicate (see [Kotlarski et al., 1981], remarks in the last paragraph on p. 292), by Enayat and Visser for base theories in purely relational language ([Enayat and Visser, 2015], remarks in Section 6), and by Leigh in the setting of this thesis (Theorem 2 in [Leigh, 2015]).

we have to show that axioms of PA are true and then by a straightforward induction we show that truth is preserved under all derivations in first-order logic. The next theorem shows that soundness of PA by itself cannot provide any new arithmetical insights.

Theorem 62 (Kotlarski–Krajewski–Lachlan, Enayat–Visser, Leigh)**.** $\mathrm{CT}^- + AS$ *is conservative over* PA.

It follows that the theory $\mathrm{CT}_{\mathrm{int}}$, i.e. CT^- with the principle of internal induction added, is also conservative, where by the internal induction principle we mean the following axiom:

$$\forall \psi(v) \in \mathrm{Form}_{\mathrm{PA}}^{\leq 1}\Big(\forall x \left(T * \psi(x) \to T * \psi(Sx)\right) \longrightarrow \left(T * \psi(0) \to \forall x\, T * \psi(x)\right)\Big). \quad \text{(INT)}$$

Thus, there exists an extremely natural conservative theory of truth, CT^-, which becomes significantly stronger than PA when augmented with Π_1 induction scheme. Moreover, we can extend CT^- in a seemingly significant way, obtaining the theory $\mathrm{CT}_{\mathrm{int}}$, which is still conservative. Now, it seems very natural to look for the "borderline cases" between CT^- and CT_1 and try to establish which properties *make compositional truth nonconservative*. In private communication, Ali Enayat proposed to dub the border between conservative and nonconservative extensions of CT^- "the Tarski's boundary." Here we adopt this term. So now we can summarize the goal of this chapter as follows: we are trying to locate the Tarski's boundary.

Let us note that there is actually one very natural theory, which can be very easily seen to be non-conservative over PA. Namely, note that the proof of the Theorem 56 actually consisted in showing the following principle:

$$\forall \phi \in \mathrm{Sent}_{\mathrm{PA}}\Big(\mathrm{Pr}_{\mathrm{PA}}(\phi) \to T\phi\Big). \quad \text{(SPA)}$$

Let us call it **the soundness principle for** PA. It is one possible form of reflection principles. It states that whatever is provable in PA is true. Since the sentence $\ulcorner 0 = 1 \urcorner$ is provably in CT^- not true, it follows that PA does not prove it. Thus, we easily conclude that CT^- + SPA is not conservative over PA.

Upon reflection on the proof of Theorem 56, it may seem that SPA is a rather unhappy mix of two very different intuitions. The first is that PA is correct, i.e., what it assumes is *true*. The other is that truth is *preserved* in first-order derivations. The latter may be argued to have much more fundamental status than the first one, which does not really belong to the realm of truth theory *per se*, but rather expresses our trust in some specific theory. These claims are of course very vague and in fact will turn out to be far from correct, but let us use them as a tentative motivation for introducing the following axiom of **reflection for first-order logic**

$$\forall \phi \in \mathrm{Sent}_{\mathrm{PA}}\Big(\mathrm{Pr}_T(\phi) \to T\phi\Big). \quad \text{(RFO)}$$

This principle states that whatever arithmetical sentence is provable from true premises in first-order logic is true. In other words, truth is closed under reasoning in first-order logic.

Now, it may be easily seen that SPA can be derived from two more basic principles, as mentioned above. Namely, it is a direct consequence of the principle of the axiomatic soundness of PA (AS) and the reflection over first order logic (RFO). Note, that adding the first principle to CT^- results in a conservative theory, whereas combining the two principles gives a theory which is immediately seen to be nonconservative.

Observe that the above considerations motivate one natural variant of reflection principles. We have considered an axiom which states that PA is sound and another one that truth is closed under first order logic. One possible new principle is the **soundness principle for first-order logic** which states that whatever arithmetical sentence is provable in pure first-order logic is true, which reads as follows:

$$\forall \phi \in \mathrm{Sent}_{\mathrm{PA}}\Big(\mathrm{Pr}_{\emptyset}(\phi) \to T\phi\Big), \tag{SFO}$$

where $\mathrm{Pr}_{\emptyset}$ means provability in the pure first-order logic.

Now, CT^-+ RFO and CT^-+ SFO seem to be good candidates to consider to pin down Tarski boundary in a more fine-grained way, as none of these principles is either immediately seen to be conservative or otherwise. We will return further to both of these theories.

Another approach which one may take to understand theories intermediate between CT_1 and CT^- is to consider even more restricted induction axioms. One obvious restriction is to take induction axioms for the extended language only for the formulae of complexity Δ_0. Let us call the theory obtained in such a way CT_0.

A version of the theory CT_0 formulated for the satisfaction predicate has been considered by Kotlarski, who has claimed that it proves SPA. Unfortunately, as observed by Richard Heck and Albert Visser,[6] Kotlarski's proof contained an essential gap. It seemed that the gap could not be overcome without Π_1-induction for the truth predicate. In joint work with Mateusz Łełyk, we have managed to show that CT_0 indeed is arithmetically no weaker than CT^-+ SPA, but using some new methods not present in Kotlarski's paper. We will present the proof of this fact in the next section.

2.2. Non-conservativity of CT_0

In this section we will present the following theorem, whose variant for the satisfaction predicate was first claimed by Kotlarski:

Theorem 63. CT_0 *is arithmetically as strong as* CT_0+ *SPA, i.e., it proves all consequences of* CT_0+ *SPA from the language* $\mathscr{L}_{\mathrm{PA}}$.

We will prove the result in a rather direct way. Let us introduce one more notion to spell out the precise relationship between CT_0 and CT_0 + SPA.

Definition 64. Let Th_1, Th_2 be two theories, let $\mathscr{L}_1$ be the language of Th_1, let $\mathscr{L}_2$ be the language of Th_2, and let $\mathscr{L}$ be some fixed language contained both in $\mathscr{L}_1$ and $\mathscr{L}_2$.

[6] To our best knowledge, these observations have never been published.

Suppose that $\mathscr{L}_2$ extends $\mathscr{L}$ only with relation symbols. We say that Th_1 **relatively defines** Th_2 over $\mathscr{L}$ if for all symbols $P_i \in \mathscr{L}_2 \setminus \mathscr{L}$, there exists a formula ϕ_i such that provably in Th_1 all the axioms of Th_2 are satisfied with formulae ϕ_i substituted for the corresponding symbols P_i of Th_2.

In other words, Th_1 **relatively defines** Th_2 if there exists interpretation of Th_2 in Th_1 which leaves $\mathscr{L}$ intact and does not restrict the quantification. Now, our result can be reformulated as follows.

Theorem 65. CT_0 *relatively interprets* CT_0 + *SPA over* PA.

In other words, we can find a formula ϕ such that provably in CT_0, this formula ϕ satisfies the compositional axioms of CT^-, Δ_0-induction scheme, and the principle of soundness for PA. This clearly entails that CT_0 + SPA is arithmetically conservative over CT_0 as observed already in [Fujimoto, 2010], where the notion of the relative definability of truth theories has been introduced.

It turns out that actually a stronger result holds. Basing on findings of Cezary Cieśliński, Mateusz Łełyk has shown that if we add to axioms of CT^- an additional technical restriction that all true objects are sentences, then CT_0 and CT^-+ SPA *are actually the same theory*. This is indeed very surprising, since the two axiomatisations do not seem to have much in common. We will return to these issues in the next section.

Let us now present a sketch of the proof of Theorem 63. The basic idea is as follows: we construct a family $(T_c(v))$ of arithmetical formulae such that all the predicates $T*T_c$ are compositional for formulae of small enough complexity for some careful choice of a complexity measure. Then we show that Δ_0-induction is actually enough to prove the internal induction principle, which by an easy application of compositional axioms holds for the predicates T_c as well. Then by Internal–External Lemma 53, the predicates $T*T_c$ are both compositional for formulae of small enough complexity and satisfy the full induction scheme. Moreover, we can define the formulae T_c so that the predicates $T*T_c$ and $T*T_d$ agree at formulae, at which both are compositional. Then we can check that the predicate defined as a "union" $\bigcup_c T*T_c$ actually satisfies compositional axioms, Δ_0-induction and SPA. This shows that in CT_0 we can relatively interpret CT_0+ SPA and thus, SPA is arithmetically no stronger than CT_0.

Now we will spell out the sketch of the above proof in a series of lemmata. We will begin with a well-known fact, which really describes how we think of Δ_0-induction for the truth predicate. Its proof may be found in [Kossak and Schmerl, 2006], Proposition 1.4.2.

Fact 66. *Let* $M \models \mathrm{PA}$ *be any model and* $A \subset M$*. Then the following conditions are equivalent:*

1. (M, A) *satisfies* Δ_0*-induction for the formulae containing the predicate* A*.*
2. *For every* $c \in M$ *the set of elements of* A *smaller than* c *is coded in* M*.*

In our thesis, we will repeatedly use the above fact without explicitly mentioning it. Now, let us state one definition and an observation, which although rather trivial, will turn to be surprisingly useful.

Definition 67. By **disjunctive correctness principle** we mean a formalised version of the following axiom:

> For all nonempty sequences $(\phi_i)_{i=0}^{c}$, if $\phi_i \in \mathrm{Sent}_{\mathrm{PA}}$ for every i and $\bigvee_{i\leq c} \phi_i$ is a disjunction of the formulae ϕ_i in the obvious order and with the parentheses grouped to the left, then

$$T\left(\bigvee_{i\leq c} \phi_i\right) \equiv \exists i \leq c\, (T\phi_i). \tag{DC}$$

In other words, disjunctive correctness states that an arbitrary finite disjunction is true if and only if one of the disjuncts is. Note that it is not obvious whether DC follows from the axioms of CT^- alone. Indeed, the compositional axioms of CT^- alone are not enough to prove disjunctive correctness. A possible counterexample to DC can be constructed as follows: for some nonstandard c, a truth predicate can render

$$\underbrace{0=1 \vee 0=1 \vee \ldots \vee 0=1}_{c \text{ times}}$$

true. Obviously, all disjuncts in the above sentence are false. Necessarily, the last disjunct, the second last disjunct and so on, would have to be false, so the sentences

$$\underbrace{0=1 \vee 0=1 \vee \ldots \vee 0=1}_{c-1 \text{ times}}$$

$$\underbrace{0=1 \vee 0=1 \vee \ldots \vee 0=1}_{c-2 \text{ times}}$$

would be true and we could not point down the first true disjunction of sentences "$0 = 1$". A truth predicate satisfying compositional axioms for which the described pathology holds may be relatively easily constructed e.g. using methods of Enayat–Visser. However, Δ_0-induction is already enough to prevent this pathology.

Lemma 68. CT_0 *proves the disjunctive correctness principle.*

Proof. We fix any sequence $(\phi_i)_{i\leq c}$ of arithmetical sentences and prove the principle for all its initial segments $(\phi_i)_{i\leq d}$ by induction on $d \leq c$. We may assume that, by convention, a disjunction of formulae from a sequence (ϕ_0) of length one is simply the formula ϕ_0. We readily check that the lemma holds in this case.

Suppose that the lemma holds for the initial segment $(\phi_i)_{i\leq d}$ of our fixed sequence and consider its initial segment $(\phi_i)_{i\leq d+1}$. Then by definition

$$\left(\bigvee_{i\leq d+1} \phi_i\right) = \left(\bigvee_{i\leq d} \phi_i\right) \vee \phi_{d+1}.$$

Which yields

$$T\left(\bigvee_{i\leq d+1}\phi_i\right)\equiv T\left(\bigvee_{i\leq d}\phi_i\right)\vee T\phi_{d+1}.$$

By induction hypothesis

$$T\left(\bigvee_{i\leq d}\phi_i\right)\equiv \exists i\leq d\,(T\phi_i).$$

Thus we conclude that

$$T\left(\bigvee_{i\leq d+1}\phi_i\right)\equiv \exists i\leq d+1\,T\phi_i.$$

□

In a similar fashion, we obtain the following lemma:

Lemma 69. CT_0 *proves the internal induction principle INT.*

Proof. Let $(M,T)\models \mathrm{CT}_0$ be an arbitrary model. Fix any $\phi(v)\in M$ such that $M\models \phi\in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}$ and an arbitrary $c\in M$. The set of elements $x\leq c$ satisfying $T(x)$ is coded. Consequently, the set of $x\leq c$ such that $(M,T)\models T*\phi(x)$ is coded as well. Then, if there is an element $x\leq c$ such that $T*\phi(x)$, then there is the least such x. Since c was arbitrary, this shows that either all elements x satisfy $T*\phi(x)$ or there is the least element which fails to satisfy this formula. This is equivalent to the induction principle for the formula $T*\phi(x)$. Since $\phi(v)$ was arbitrary, we have proved the internal induction principle. □

Now we will introduce the main tool of our proof: some carefully chosen family of arithmetical truth predicates T_c which will yield partially compositional inductive truth predicates as described in the above sketch.

Definition 70. We define A_n as a class of those formulae, whose syntactic tree has height n. More precisely, we define them by induction as follows:

$$\phi\in A_0\equiv \exists s,t\in \mathrm{Term}_{\mathrm{PA}}\ \phi=(s=t)$$

$$\phi\in A_{n+1}\equiv\begin{cases} & \exists\psi\in A_n & \phi=(\neg\psi) & \\ \vee & \exists k,l\,\exists\psi\in A_k,\eta\in A_l & \phi=(\psi\wedge\eta) & \wedge\max(k,l)=n\\ \vee & \exists k,l\,\exists\psi\in A_k,\eta\in A_l & \phi=(\psi\vee\eta) & \wedge\max(k,l)=n\\ \vee & \exists\psi\in A_n\exists v & \phi=(\exists v\ \psi) & \\ \vee & \exists\psi\in A_n\exists v & \phi=(\forall v\ \psi). & \end{cases}$$

The definition of the classes A_n is primitive recursive, so it may be represented with an arithmetical formula $x\in A_y$ with two free variables x,y. Having defined the classes A_n, we may introduce truth predicates which work smoothly for formulae in some fixed class A_n. We begin with one auxiliary definition.

Definition 71. We define formulae Θ_n by induction on n as follows:

$$\Theta_0(\phi) \equiv \exists s,t \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(\phi = (s=t) \wedge s^\circ = t^\circ\Big).$$

$$\Theta_{n+1}(\phi) \equiv \begin{cases} & \exists \psi & \phi = (\neg\psi) \wedge \neg\Theta_n(\phi) \\ \vee & \bigvee_{k,l\leq n} \exists \psi \in A_{\underline{k}}, \eta \in A_{\underline{l}} & \phi = (\psi \wedge \eta) \wedge \Theta_k(\psi) \wedge \Theta_l(\eta) \\ \vee & \bigvee_{k,l\leq n} \exists \psi \in A_{\underline{k}}, \eta \in A_{\underline{l}} & \phi = (\psi \vee \eta) \wedge \big(\Theta_k(\psi) \vee \Theta_l(\eta)\big) \\ \vee & \exists \psi, v & \phi = (\exists v\ \psi) \wedge \big(\exists x \Theta_n(\psi(\underline{x}))\big) \\ \vee & \exists \psi, v & \phi = (\forall v\ \psi) \wedge \big(\forall x \Theta_n(\psi(\underline{x}))\big) \end{cases}$$

Note that the definition of the predicates Θ is quite similar to the definition of the usual truth predicates for the classes Σ_n. It keeps track of the syntactic build-up of the formulae which it is applied to and inductively unravels these syntactic structures.

Since the definition of the formulae Θ_n is primitive recursive, it can be formalised in PA, so it makes sense to speak of formulae Θ_c for nonstandard c which allows us to introduce the main tool of this section.

Definition 72. We define a family (T_c) of arithmetical truth predicates as follows:

$$T_c(x) = \bigvee_{i=0}^{c} x \in A_{\underline{i}} \wedge \Theta_i(x).$$

Note that in the above definition the numeral $\underline{i}$ occurs only once. Namely, the formula $x \in A_{\underline{i}}$ is an actual standard binary formula with a possibly nonstandard term substituted for one of the variables. The formula $\Theta_i(x)$ cannot be written down in such a form, so Θ_i-s are actually different formulae as the parameter i varies.

Now we are in the position to show the main feature of the formulae T_c. They give rise to truth predicates which are compositional for the formulae in a fixed class A_n.

Lemma 73. *Let $(M,T) \models \mathrm{CT}_0$ and let $c \in M$ be an arbitrary element. Then for any $j \leq c$ the formula $T * T_c$ is compositional at $A_j = \{x \in M \mid M \models x \in A_j\}$.*

Recall that a formula is compositional at some subset of M, if it satisfies compositional axioms of CT^- for all (codes of) sentences in that set, see Definition 48.

Proof of Lemma 73. Fix any $(M,T) \models \mathrm{CT}_0$, any $c \in M$, and an arbitrary $j \leq c$. We have to check that for any sentence $\phi \in A_j$ the formula $T * T_c$ is compositional at ϕ. We check it by cases depending on the main connective or quantifier in ϕ. Let us consider three cases: ϕ is atomic, ϕ is a conjunction or ϕ is an existential formula.

(I) If $\phi = (s=t)$ for some $s,t \in \mathrm{ClTerm}_{\mathrm{PA}}$, then the following chain of equivalences

holds:

$$\begin{aligned} T * T_c(s=t) &= T\Big(\bigvee_{j=0}^{c} \underline{s=t} \in A_{\underline{j}} \wedge \Theta_j(\underline{s=t})\Big) \\ &\equiv \exists j \leq c\, T\Big(\underline{s=t} \in A_{\underline{j}} \wedge \Theta_j(\underline{s=t})\Big) \\ &\equiv \exists j \leq c\, (s=t) \in A_j \wedge T(\Theta_j(\underline{s=t})) \\ &\equiv T(\Theta_0(\underline{s=t})) \\ &= T\Big(\exists p, q \in \mathrm{ClTerm}_{\mathrm{PA}}\, \Big(\underline{s=t} = (p=q) \wedge p^\circ = q^\circ\Big)\Big) \\ &\equiv s^\circ = t^\circ. \end{aligned}$$

The first equivalence holds by disjunctive correctness, the second follows by Proposition 51 due to the fact that $x \in A_y$ is a standard binary formula. The third holds, because provably in PA any atomic formula belongs to A_0 and to no other class A_j. The last equivalence holds, since there is exactly one pair of (codes of) closed terms, which can form the atomic sentence ϕ. This proves the case where ϕ is an atomic sentence.

(II) If $\phi = \psi \wedge \eta$ for some $\psi, \eta \in \mathrm{Sent}_{\mathrm{PA}}$, then if $\phi \in A_j$ for some $j \leq c$, then actually $\psi \in A_k, \eta \in A_l$ for some $k, l < j$ and the following chain of equivalences holds:

$$\begin{aligned} T * T_c(\psi \wedge \eta) &= T\Big(\bigvee_{j=0}^{c} \underline{\psi \wedge \eta} \in A_{\underline{j}} \wedge \Theta_j(\underline{\psi \wedge \eta})\Big) \\ &\equiv \exists j \leq c\, (\psi \wedge \eta) \in A_j \wedge T * \Theta_j(\underline{\psi \wedge \eta}). \end{aligned}$$

Since

$$M \models \exists k, l \leq j\, \exists \psi \in A_k, \eta \in A_l\, (\phi = \psi \wedge \eta),$$

the last condition in the above chain holds if and only if

$$T * \Theta_k(\psi) \wedge T * \Theta_l(\eta).$$

Now, again we observe that k, l are unique x, y such that $\psi \in A_x$ and $\eta \in A_y$, which allows us to draw the following conclusion:

$$T * T_c(\psi) \wedge T * T_c(\eta).$$

This concludes the proof in the case, when ϕ is a conjunction.

(III) Suppose that $\phi \in A_{j+1}$ is an existential sentence, i.e., there exists a (code of a) formula $\psi(v) \in A_j$ with at most one free variable such that $\phi = \exists v\, \psi(v)$. Then the following chain of equivalences holds:

$$\begin{aligned} T * T_c(\exists v \psi(v)) &\equiv T * \Theta_{j+1}(\exists v \psi(v)) \\ &\equiv \exists x\, T * \Theta_j(\psi(\underline{x})) \\ &\equiv \exists x\, T * T_c(\psi(\underline{x})). \end{aligned}$$

As in the previous cases, the second step follows by the uniqueness of the syntactic structure of formulae and the last step follows by disjunctive correctness.

The other cases are similar and we omit them. □

Now, let $(M,T) \models \mathrm{CT}_0$. Take any standard formula $\phi \in \mathscr{L}_{\mathrm{PA}P}$.[7] Note that, since for any c, $T_c \in \mathrm{Form}_{\mathrm{PA}}$, the following holds:

$$(M,T) \models \forall x \left(T * \phi[T_c](x) \to T * \phi[T_c](Sx)\right) \longrightarrow \left(T * \phi[T_c](0) \to \forall x\, T * \phi[T_c](x)\right).$$

The predicate T is compositional at the whole $\mathrm{Sent}_{\mathrm{PA}}$, so as a corollary of Internal–External Lemma 53, we conclude that:

$$(M,T) \models \forall x \left(\phi[T * T_c](x) \to \phi[T * T_c](Sx)\right) \longrightarrow \left(\phi[T * T_c](0) \to \forall x\, \phi[T * T_c](x)\right).$$

Since ϕ is an arbitrary (semirelational) formula with at most one free variable, we obtain the following lemma:

Lemma 74. *Let $(M,T) \models \mathrm{CT}_0$ and let $c \in M$ be an arbitrary element. Then the formula $T * T_c(x)$ satisfies the full induction scheme.*

The truth predicate in CT enjoys a number of good properties such that their proof actually does not require full compositionality, but rather compositionality for subformulae of some fixed formula or a fixed coded set of formulae. Once we have proved that the predicates $T * T_c$ are fully inductive, we may show that they enjoy many of these properties.

Lemma 75 (Extensionality for $T * T_c$)**.** *Let $(M,T) \models \mathrm{CT}_0$. Then for each a there exists some $b \in M$ such that the formula $T * T_b$ satisfies the extensionality principle for formulae no greater than a, i.e.,*

$$\forall \phi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}, \phi \leq a\, \forall s,t \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(s^\circ = t^\circ \to T * T_b\phi(s) \equiv T * T_b\phi(t)\Big).$$

Lemma 76 (RFO for $T * T_c$)**.** *Let $(M,T) \models \mathrm{CT}_0$ and let $d, \phi, c \in M$. Suppose that*

$$(M,T) \models \mathrm{Prov}_{T*T_c}(d,\phi)$$

and that all formulae in the proof d are in fact in A_c. Then

$$(M,T) \models T * T_c(\phi).$$

Recall that $\mathrm{Prov}_{T*T_c}(d,\phi)$ means that d is a proof of ϕ in sequent calculus such that for all initial sequents $\longrightarrow \eta$, the model (M,T) satisfies $T * T_c(\eta')$ for all sentences η' resulting from substituting closed terms for eigenvariables in the formula η, see Definition 6.

Lemma 77 (SPA for $T * T_c$)**.** *Let $(M,T) \models \mathrm{CT}_0$ and let $d, \phi, c \in M$. Suppose that*

$$(M,T) \models \mathrm{Prov}_{\mathrm{PA}}(d,\phi)$$

and that all formulae occurring in the proof d are in A_c. Then

$$(M,T) \models T * T_c(\phi).$$

[7] For the definition of $\mathscr{L}_{\mathrm{PA}P}$ and the notation $\phi[T_c]$, see Definition 43.

The proofs of the above Lemmata are completely analogous to the proofs of EXT, SPA and RFO in CT, using the fact that in those arguments, we only make use of compositionality *at the formulae occurring in a given derivation* d in sequent calculus. Except for rather awkward formulation, both Lemmata are straightforward adaptations of the formerly proved facts.

Note, that Lemma 77 already implies that CT_0 is not conservative over PA. Namely, suppose that M is a model of PA in which there is a proof d in PA of the sentence $0 = 1$. If (M', T) is a model of CT_0 such that M is an elementary submodel of M', then by our lemma $T * T_c(0 = 1)$ would hold for some large enough c, which in turn would contradict the compositionality Lemma 73. Let us conclude this section by presenting the results obtained so far in somewhat cleaner form.

Definition 78. Let us define the following predicate in CT_0:

$$T'(x) = \exists c\, T * T_c(x).$$

In order for this definition to work smoothly, we have first to make sure that the predicates $T * T_c$ are actually compatible as c varies.

Lemma 79. *Let $(M, T) \models \mathrm{CT}_0$ be an arbitrary model. Let $c < d$ be arbitrary two elements. Suppose that $M \models \phi \in A_e$ for some $e \leq c$. Then*

$$(M, T) \models T * T_c(\phi) \equiv T * T_d(\phi).$$

Proof. Let $(M, T), c, d, e, \phi$ be as above. Then by disjunctive correctness, definition of the formulae T_c, T_d and the fact that e is the unique element such that

$$M \models \phi \in A_e,$$

we may conclude that

$$(M, T) \models T * T_c(\phi) \equiv T * \Theta_e(\phi) \equiv T * T_d(\phi).$$

□

The above lemma guarantees that T' is reasonably well-behaved.

Lemma 80 (Compositionality of T')**.** *Let $(M, T) \models \mathrm{CT}_0$. Then*

$$(M, T') \models \mathrm{CT}^-.$$

Proof. Let $(M, T) \models \mathrm{CT}_0$. By Lemma 79, for an arbitrary $c < d \in M$, $\phi \in \mathrm{Sent}_{\mathrm{PA}} \cap A_c$ the following holds:

$$(M, T) \models T'(\phi) \equiv T * T_d(\phi).$$

The claim follows by Lemma 73. □

We have already checked that in any model $(M, T) \models \mathrm{CT}_0$, the formula T' defined as above satisfies CT^-. We will show that it actually satisfies full CT_0.

Lemma 81. *Let* $(M, T) \models \mathrm{CT}_0$. *Then* $(M, T') \models \mathrm{CT}_0$ *as well.*

Proof. Fix any $(M, T) \models \mathrm{CT}_0$. By Fact 66, it is enough to check whether for any $c \in M$ the set $[0, c] \cap T'$ is coded. But by Lemma 79

$$[0, c] \cap T' = [0, c] \cap T * T_c,$$

and the latter set is clearly coded, since $(M, T * T_c)$ satisfies the full induction scheme. □

Finally, we will check that T' satisfies the reflection principles which we have defined previously.

Lemma 82 (Reflection principles for T')**.** *Let* $(M, T) \models \mathrm{CT}_0$. *Then* (M, T') *satisfies RFO and SPA.*

Proof. Let $(M, T) \models \mathrm{CT}_0$. We will show that T' satisfies SPA. Pick any $d, \phi \in M$ such that

$$(M, T) \models \mathrm{Pr}_{\mathrm{PA}}(d, \phi).$$

Then by Lemma 79, choosing c such that all formulae in the proof d are in the class A_c,

$$(M, T) \models T'(\phi) \equiv T * T_c(\phi),$$

but, by Lemma 77,

$$(M, T) \models T * T_c(\phi),$$

so indeed

$$(M, T) \models T'(\phi).$$

□

This concludes our proof. The formula T' defines a relative interpretation of CT_0 with additional reflection principles within CT_0. This proves Theorem 65 and (immediately) gives the following corollary:

Corollary 83. CT_0 *relatively interprets* CT_0*+ RFO + SPA.*

Note that all we have used in our proof was INT and DC. We have only used Δ_0-induction for the truth predicate to show that these principles hold. Thus, we actually obtain a stronger corollary.

Corollary 84. CT^-*+ DC + INT relatively interprets* CT_0*+ RFO + SPA.*

2.3. Tarski's boundary

In the previous section, we have shown that a surprisingly innocent-looking theory $\mathrm{CT}^- + \mathrm{DC} + \mathrm{INT}$ actually relatively interprets strong reflection principles. It is even more striking as one realises that the theory $\mathrm{CT}^- + \mathrm{INT}$ is conservative over PA (see Theorem 62). Hence the principle that somewhat resembles reflection principles is by itself on the "weak" side of the Tarski's boundary. It is only after adding some amount of generalised compositionality in the form of DC, that this theory gets some actual boost.

From the interpretability results in the previous section, it follows that the arithmetical consequences of CT_0, $\mathrm{CT}_0 + \mathrm{SPA}$ and $\mathrm{CT}^- + \mathrm{DC} + \mathrm{INT}$ are simply the same. This is not the first surprising result characterising CT_0. In [Cieśliński, 2017], Theorem 12.3.1, Cieśliński has shown that CT_0 is *the same theory* as CT^- enriched with the following technical assumption that only sentences are true, which we call the **normality principle:**

$$\forall x \Big(T(x) \to \mathrm{Sent}_{\mathrm{PA}}(x) \Big), \tag{NORM}$$

and with the following principle of **propositional reflection:**[8]

$$\forall \phi \in \mathrm{Sent}_{\mathrm{PA}} \Big(\mathrm{Pr}^{\mathrm{pr}}_T(\phi) \to T\phi \Big), \tag{RP}$$

where $\mathrm{Pr}^{\mathrm{pr}}$ denotes provability in classical propositional logic and the conventions governing the subscript are the same as in the case of the first-order provability predicate Pr. Thus, $\mathrm{Pr}^{\mathrm{pr}}_T$ denotes provability in classical propositional logic from true premises and the principle RP states that truth is closed under derivations in that logic.

The theory $\mathrm{CT}^- + \mathrm{RP}$ looks quite innocent and actually we have spent over a year trying to show in collaboration with Cezary Cieśliński and Mateusz Łełyk that CT_0 is conservative over PA, precisely by showing that any model M of PA may be elementarily extended to a model carrying a truth predicate satisfying $\mathrm{CT}^- + \mathrm{RP}$.

Another important connection known prior to the results presented in this paper was discovered also by Cieśliński, who showed that both the principle of propositional reflection RP and the principle of soundness for first-order logic entail internal induction INT (see [Cieśliński, 2010a], Theorem 4 and [Cieśliński, 2010b], Theorem 1, respectively for the two results). This is an extremely nice findings. Namely, RP and SFO seem to be principles of purely logical nature, since they just describe how logic and truth interact. On the other hand, the internal induction principle states that some arithmetical principle is true. Vaguely speaking, it seems to express our trust in the correctness of the induction scheme, rather than some very basic logical principle. Cieśliński's result shows that those vague intuitions are not accurate, since we are in the position to justify internal induction purely on the basis of reflection-like axioms.

[8] An analogous result has been claimed in [Cieśliński, 2010a], Theorem 4, for $\mathrm{CT}^- + \mathrm{RP}$ without the technical axiom that only sentences are true. It turned out that the proof not employing this additional condition contains a gap.

Once we learned that CT_0, CT^-+ RP, CT^-+ DC + INT, CT_0+ RFO and CT_0+ RFO + SPA share arithmetical consequences, we still supposed that these theories were not actually the same. They were conjectured to have different truth-theoretic consequences, as they were formed by appealing to very different intuitions concerning the truth predicate. Moreover, there were some other truth-theoretic principles, whose status was not clear and which seemed to be reasonable candidates for principles of the intermediate strength between full CT_0 and conservative theories like CT^-.

Finally, Cezary Cieśliński, Ali Enayat, and Mateusz Łełyk combining some new results and some previously discovered facts, both published and not, managed to reach the following neat theorem. It shows that CT^-+ SPA, a theory which is automatically seen to be nonconservative over PA, is actually very robust and admits a number of different and seemingly remote characterisations upon adding the normality principle to the list of axioms of CT^-.

Theorem 85 (Cieśliński–Enayat–Łełyk)**.** *The following axioms yield the same theory:*

1. CT_0+ *NORM.*
2. CT_0+ *NORM* + *RFO* + *AS.*
3. CT^-+ *NORM* + *RFO.*
4. CT^-+ *NORM* + *SPA.*
5. CT^-+ *NORM* + *SFO.*
6. CT^-+ *NORM* + *RP.*
7. CT^-+ *NORM* + *DC* + *AS.*

Let us repeat once more: all listed systems of axioms yield exactly the same theory, not only theories which share arithmetical consequences or relatively interpret each other. They all have the same arithmetical consequences as CT_0. As observed by Cieśliński, the inclusion of the additional axiom is necessary for the strict equality to be true. For example, take any nonstandard model $(M, T) \models \mathrm{CT}^-$+ NORM + SPA. Then add all the standard numbers to the extension of the truth predicate. The resulting model still satisfies CT^-+ SPA (since the axioms of that theory only enforce certain behaviour of the truth predicate on the set $\mathrm{Sent}_{\mathrm{PA}}(M)$), but obviously it cannot satisfy CT_0.

Recall that previous research has revealed that CT^-+ AS is still conservative over PA. Thus we encountered an extremely interesting phenomenon: a number of natural truth theories turned out to be either conservative or to be *precisely* the obviously nonconservative theory CT^-+ SPA.

There are still a couple of natural theories whose status is not quite clear to us. Let us close this section by listing theories for which syntactic conservativity over PA is still an open problem. The first principle missing in our list is the **propositional soundness principle**:

$$\forall \phi \in \mathrm{Sent}_{\mathrm{PA}} \Big(\mathrm{Pr}^{\mathrm{pr}}_{\emptyset}(\phi) \longrightarrow T\phi \Big). \tag{SP}$$

It states that any arithmetical sentence provable in pure propositional logic is true. Thus, it resembles both the propositional reflection principle RP, which says that truth is preserved under reasoning in classical propositional logic, and first-order soundness principle SFO which states that any sentence *provable* in pure first-order logic is true. The principle SP stating simply that tautologies of propositional logic are true may seem extremely weak, but it is very similar to other principles which we also conjectured to give conservative theories of truth.

The other theory which naturally emerges in our research is $\mathrm{CT}^- + \mathrm{DC}$, the theory of compositional truth with the disjunctive correctness property. Again, this principle seems very weak, as it is simply a form of generalised compositionality. On the other hand, we known that this very principle coupled with a weak axiom of internal induction, yields a surprisingly strong theory. Thus we are left with the following questions:

Problem 86. Are the following theories proof-theoretically conservative over PA?

1. $\mathrm{CT}^- + \mathrm{DC}$.[9]
2. $\mathrm{CT}^- + \mathrm{SP}$.

As for the present moment, we are in two minds what is the expected answer for these questions.

[9]Recently, Fedor Pakhomov has solved this problem by showing that $\mathrm{CT}^- + \mathrm{DC}$ is arithmetically as strong as CT_0.

Chapter 3

Model-theoretic strength I: classical theories

In this chapter, we will turn to a more fine-grained way of comparing axiomatic truth theories. Proof-theoretic conservativity, although very natural, seems to be somewhat rough way of comparing theories of truth. For example, disquotational theories, like TB and UTB and the pure compositional theory of truth CT^{-} are all conservative over PA, although they still seem very different. Model-theoretic considerations provide a good tool for expressing this difference. Let us introduce the key notion of this chapter.

Definition 87. Let $\mathrm{Th} \subseteq \mathrm{Th}'$ be any two theories and let $\mathscr{L}'$ be the language of the theory Th'. We say that Th' is **semantically (model-theoretically) conservative** over Th if for any model $M \models \mathrm{Th}$ there exists an expansion $(M, P_1, \ldots, P_\alpha)$ to a model of the theory Th'.

In our case, Th will be Peano Arithmetic. Hence a theory of truth Th is semantically conservative over PA if in every model M of PA, there is a subset $T \subset M$ such that (M, T) forms a model of Th. In still other words, in every model one can find a set of (codes of) sentences satisfying the conditions which Th puts on the truth predicate.

In analogy to the case of syntactic conservativity, along with the notion which simply divides theories in two categories: strong and weak, there comes a method of comparing strength of theories.

Definition 88. Let us fix some theory B in a language $\mathscr{L}$ and let $\mathrm{Th}, \mathrm{Th}'$ be any two theories in languages extending $\mathscr{L}$. We say that Th' is **semantically no weaker than** Th over B, if every model M of B which can be expanded to a model of Th', can be also expanded to a model of Th. We denote this relation by

$$\mathrm{Th} \leq_{\mathrm{mod,B}} \mathrm{Th}'.$$

We will also say that Th' is **semantically strictly stronger** than Th or that both theories **have the same strength** defining these notions in the obvious way and denoting them with $<_{\mathrm{mod,B}}, =_{\mathrm{mod,B}}$. In our thesis, we focus on the case where $\mathrm{B} = \mathrm{PA}$, so whenever we use expressions like "semantically stronger", "semantically weaker" without

mentioning the theory B, we implicitly assume that B $=$ PA. Similarly, when we use the symbols $\leq_{\text{mod}}, <_{\text{mod}}$ etc., we mean $\leq_{\text{mod,PA}}, <_{\mathscr{L}_{\text{PA}}}$ etc.

Actually in this chapter, we will simply write $\leq, \geq, <, >, =$ to denote the relations of model-theoretic strength, since there will be no risk of confusion with the notation introduced in the previous chapter or with any other ordering we will consider.

3.1. Models of disquotational truth

In this section, we discuss the semantic strength of the classical disquotational truth theories. It turns out that already a theory as weak as TB is not model-theoretically conservative over PA. Let us recall that the truth-theoretic axioms of the former comprise only Tarski's biconditionals for arithmetical sentences and the full induction scheme.

Let us begin with a rather trivial result.

Proposition 89. UTB$^-$ *(and consequently* TB$^-$*) is semantically conservative over* PA*.*

Proof. Let M be any model of PA. Let $T \subset M$ be defined in the following way:

$$\{\phi(\bar{t}) \in M \mid \phi(v_1, \ldots, v_n) \in \text{Form}_{\text{PA}} \cap \omega \wedge \bar{t} \in \text{ClTermSeq}_{\text{PA}}(M) \wedge M \models \phi(\overline{t^\circ})\}.$$

Basically, T is the elementary diagram of M. The only difference is that we use (the codes of) terms rather than parametres (which would not quite make sense in our context).

That $(M, T) \models \text{UTB}^-$ follows directly be the definition of T. □

It turns out that once we add full induction for the truth predicate, even to the basic Tarski's scheme, the resulting theory is not semantically conservative any more. This result is due to Cieśliński and (independently) Engström[1].

Proposition 90 (Cieśliński, Engström)**.** TB *is not semantically conservative over* PA*.*

Proof. Let (M, T) be any nonstandard model of TB such that $\text{Th}(M) \neq \text{Th}(\mathbb{N})$, i.e., M and $\mathbb{N}$ do not satisfy the same arithmetical sentences. Let us consider the following type p:

$$p(x) = \{\ulcorner \phi \urcorner \in x \equiv \phi \mid \phi \in \text{Sent}_{\text{PA}} \cap \omega\}.$$

If an element c realises p, then it codes the theory of M. One can check that p is indeed realised in M by considering the following formula:

$$\exists x \forall y \leq b \Big(y \in x \equiv y \in \text{Sent}_{\text{PA}} \wedge T(y) \Big).$$

This formula is clearly satisfied for all $b \in \omega$. Therefore, by a simple overspill argument, it is satisfied by some nonstandard $b' \in M$. Let us then take any a such that

$$(M, T) \models \forall y \leq b' \Big(y \in a \equiv T(y) \Big).$$

[1] See [Cieśliński, 2015b], Theorem 7.

By disquotation scheme, it follows that a realises p in M.

We claim that the type p is not realised in an arbitrary model $K \models \mathrm{PA}$. Let $K = K(M, \emptyset)$ be the substructure of M consisting of elements definable without parametres. Since $\mathrm{Th}(M) \neq \mathrm{Th}(\mathbb{N})$, there are nonstandard definable elements, so $K \neq \mathbb{N}$. By Fact 38, we actually have $K \preceq M$. We claim that K omits the type p.

Suppose that the type p is realised by some element $b \in K$. Since this element is definable in K, there is some formula $\theta \in \mathscr{L}_{\mathrm{PA}}$ such that:

$$K \models \forall x \Big(\theta(x) \equiv x = b\Big).$$

Since K and M have precisely the same theory, for all arithmetical sentences ϕ the following holds:

$$K \models \phi \equiv \forall x \Big(\theta(x) \to \ulcorner \phi \urcorner \in x\Big).$$

But this is the arithmetical definition of truth for the theory $\mathrm{Th}(K)$, which contradicts Tarski's Theorem 15. □

Actually, from the proof of the above proposition, a characterisation of models of TB may be extracted.

Theorem 91 (Cieśliński, Engström)**.** *The following conditions are equivalent for nonstandard models $M \models \mathrm{PA}$:*

1. *M realises the type $p(x) = \{\ulcorner \phi \urcorner \in x \equiv \phi \mid \phi \in \mathrm{Sent}_{\mathrm{PA}} \cap \omega\}$.*
2. *M expands to a model of* TB.

Proof. The implication (1.) $\to$ (2.) has been shown in the proof of Proposition 90. Let us prove the converse implication. Pick a nonstandard model $M \models \mathrm{PA}$ and suppose that the element $c \in M$ realises p. Then let

$$T = \{x \in M \mid x \in c\},$$

i.e., it is the extension of the element c, viewed as a coded set. By definition of p, it follows that for any $\ulcorner \phi \urcorner \in \omega$ the equivalences

$$T(\ulcorner \phi \urcorner) \equiv \ulcorner \phi \urcorner \in c \equiv \phi$$

hold. Moreover, T obviously satisfies the full induction scheme, since it is arithmetically definable in M (with a parameter). □

Along the similar lines, one can obtain a result on the strength of UTB.[2]

Proposition 92. *Let (M, T) be any nonstandard model of* UTB*. Then M is recursively saturated.*

[2]The result is well-known in the literature. E.g., the presented argument essentially appears in [Kotlarski, 1991], in a comment following Theorem 3, formulated for CT_1, i.e. compositional truth theory with Σ_1-induction, but with a proof which does not really uses compositionality.

Proof. Let $(M,T) \models$ UTB be any nonstandard model. Pick a recursive type $p = = \{\phi_i(x, a_1, \ldots, a_n) \mid i \in \omega\}$, where $a_1, \ldots, a_n$ are some fixed elements of M. Let $(\phi_i)_{i=0}^{\infty}$ be any recursive enumeration of this type.

Now, since p is finitely realised in M, the following holds for any $b \in \omega$:

$$(M,T) \models \exists x \forall i \leq b \, T\phi_i(\underline{x}, \underline{a_1}, \ldots, \underline{a_n}).$$

Hence by overspill, there exists a nonstandard b' such that

$$(M,T) \models \exists x \forall i \leq b' \, T\phi_i(\underline{x}, \underline{a_1}, \ldots, \underline{a_n}).$$

In particular, there exists some $c \in M$ such that for any $i \in \omega$,

$$(M,T) \models T\phi_i(\underline{c}, \underline{a_1}, \ldots, \underline{a_n}).$$

Since (M,T) satisfies the uniform diquotation scheme, this implies that c realises the type p in M. □

Proposition 92 may not be reversed, because of the following (difficult) theorem:

Theorem 93 (Kaufmann–Shelah)**.** *There exists a model $M \models$ PA, such that M is recursively saturated and for every set $A \subset M$, if for every $c \in M$ the set $A \cap [0,c]$ is coded in M, then A is definable in M (with parametres).*

Models whose all piecewise coded subsets are actually definable are called **rather classless**. The proof of the above theorem under additional set-theoretic assumption $\diamond$ has been given in [Kaufmann, 1977]. Subsequently, Shelah has shown that the assumption may be dropped (see [Shelah, 1978], Application C, p. 74). A proof purely in ZFC may be found in [Schmerl, 1981], Theorem 6.

Corollary 94. *There exists a recursively saturated model $M \models$ PA which does not expand to a model of* UTB.

Proof. Let M be rather classless recursively saturated model of PA. Suppose that $(M,T) \models$ UTB for some $T \subset M$. Since (M,T) satisfies the full induction scheme, every subset of the form $T \cap [0,a]$ is coded in M. But T cannot be definable in M, since that would contradict Tarski's Theorem 15. □

One may wonder, whether some form of the above reasoning could not be carried out in TB as well. This is not the case, as we will show in the next proposition. A version of this result has occurred in our Master's thesis and in [Łełyk and Wcisło, 2017a], Theorem 3.7. This particularly simple argument is due to Albert Visser.

Theorem 95. *There exists a nonstandard model $(M,T) \models$ TB with M not short recursively saturated.*

Proof. Let $(M, T) \models \mathrm{TB}$ be any nonstandard model. As before, let us define a type p in the following way:

$$p(x) = \{\ulcorner \phi \urcorner \in c \equiv \phi \mid \ulcorner \phi \urcorner \in \mathrm{Sent}_{\mathrm{PA}} \cap \omega\}.$$

Let us fix any $a \in M$ realising this type, i.e. any code of the theory of M. Let

$$K = K(M, a)$$

be the submodel of elements definable in M with the parameter a. By Fact 38, K is an elementary submodel of M. In particular $\mathrm{Th}(M) = \mathrm{Th}(K)$. Since K contains the code of $\mathrm{Th}(M) = \mathrm{Th}(K)$, it may be expanded to a model of TB. On the other hand, by Proposition 41, K is not short recursively saturated. □

3.2. Disjunctions with stopping conditions

In our model-theoretic considerations, we will make repeated use of a certain construction in propositional logic. Since the construction is rather intricate, let us first begin with a motivating example. Suppose (M, T) is a nonstandard model of CT^-. As we have seen in the previous chapter, nonstandard arithmetical truth predicates may be a tool of certain interest. One could naïvely hope that the following arithmetical partial truth definition (where Tr_j denotes the arithmetical truth predicate for the sentences in the class Σ_j):

$$\Theta(x) = \bigvee_{j=0}^{c} x \in \mathrm{Sent}_{\mathrm{PA}} \wedge x \in \Sigma_j \wedge \mathrm{Tr}_j(x),$$

would yield in presence of a compositional truth predicate a partial truth predicate which would satisfy Tarski's biconditionals whenever c is nonstandard. More precisely, one could hope that for standard sentences ϕ, say $\phi \in \Sigma_n$, we would have

$$\begin{aligned} T * \Theta(\phi) &\equiv \exists j \leq c \left(\phi \in \Sigma_j \wedge T * \mathrm{Tr}_j(\phi)\right) \\ &\equiv T * \mathrm{Tr}_n(\phi) \\ &\equiv \phi. \end{aligned}$$

Unfortunately, the first step in the above argument is obviously wrong. Basically, it makes use of disjunctive correctness, whereas in this chapter we will be mostly preoccupied with weak theories of truth in which we cannot hope DC to hold.

It turns out that there is in fact a method to overcome this difficulty. Suppose we are given two nonstandard coded sequences of formulae (α_i), (β_i). Then we can define a (nonstandard) formula S which "behaves" according the following instructions:

- Find the first i_0 such that α_{i_0} is true.
- Then the whole formula is true if β_{i_0} is true and false if β_{i_0} is false.

The definition will actually allow that α_i and β_i are formulae rather than sentences, so different i_0 will be chosen for different elements. For example, we could choose $\alpha_i(x)$ to be "$x \in \Sigma_i \wedge x \in \mathrm{Sent}_{\mathrm{PA}}$" and β_i to be "$\mathrm{Tr}_i(x)$." Thus we could circumvent the use of disjunctive correctness and define an arithmetical truth predicate which would work for all standard sentences.

Let us now define the formulae we tried to motivate. When looking at the definition, it is probably best to keep in mind what our formula is supposed to do, since the definition basically consists in writing these instructions down.

Definition 96. Let $\alpha = (\alpha_i)_{i=0}^c, \beta = (\beta_i)_{i=0}^c$ be any two sequences of sentences. Then we define the **disjunction of β_i's with a stopping condition** α, denoted

$$\bigvee_{i=k}^{c,\alpha} \beta_i,$$

by backward induction on k:

$$\bigvee_{i=c}^{c,\alpha} \beta_i = \alpha_c \wedge \beta_c,$$

$$\bigvee_{i=k}^{c,\alpha} \beta_i = \neg(\alpha_k \wedge \neg\beta_k) \wedge \left((\alpha_k \wedge \beta_k) \vee \bigvee_{i=k+1}^{c,\alpha} \beta_i \right).$$

Note that the above definition really just mimics the informal instruction for the behaviour of the disjunctions with stopping condition which we have sketched above. The next Lemma basically says that it works fine already in PT^-.

Lemma 97. *Let (M, T) be any model of PT^- and let c be an arbitrary element of M. Let $\alpha = (\alpha_i)_{i=0}^c, \beta = (\beta_i)_{i=0}^c$ be arbitrary sequences of sentences coded in M. Suppose that for any standard k, the model (M, T) satisfies $T\alpha_k \equiv \neg T\neg\alpha_k$ and that the first k_0 such that $T\alpha_{k_0}$ holds is standard. Then*

- $T \bigvee_{i=0}^{c,\alpha} \beta_i \equiv T\beta_{k_0}$,
- $T\neg \bigvee_{i=0}^{c,\alpha} \beta_i \equiv T\neg\beta_{k_0}$.

Note that, despite this rather lengthy formulation, Lemma 97 really spells out that we can define nonstandard formulae whose truth will depend just on the truth of β_{i_0}, where i_0 is defined as the number, at which some condition α is met. Crucially, the truth of the formula will not depend on what happens in the "tail" of the disjunction.

Proof of Lemma 97. Let α, β, k_0 be as in the assumptions of the lemma. We prove by backward induction on $k \leq k_0$ that

- $T \bigvee_{i=k}^{c,\alpha} \beta_i \equiv T\beta_{k_0}$,
- $T\neg \bigvee_{i=k}^{c,\alpha} \beta_i \equiv T\neg\beta_{k_0}$.

Let us focus on the first item, the second being fully analogous. First suppose that $k = k_0$. If $k_0 = c$, then the claim follows by the assumption that either $T\alpha_j$ or $T\neg\alpha_j$ holds for any standard j. If $k_0 < c$, then

$$T \bigvee_{i=k}^{c,\alpha} \beta_i \equiv T\neg(\alpha_k \wedge \neg\beta_k) \wedge T\left((\alpha_k \wedge \beta_k) \vee \bigvee_{i=k+1}^{c,\alpha} \beta_i\right)$$

which is equivalent to the following:

$$(T\neg\alpha_k \vee T\beta_k) \wedge \left((T\alpha_k \wedge T\beta_k) \vee T \bigvee_{i=k+1}^{c,\alpha} \beta_i\right).$$

By assumption, $T\alpha_{k_0}$ holds and (by assumption on formulae α_i) $T\neg\alpha_{k_0}$ does not hold. Thus, by propositional logic the formula $T\neg\alpha_{k_0} \vee T\beta_{k_0}$ is equivalent to $T\beta_{k_0}$ and the formula $\left((T\alpha_k \wedge T\beta_k) \vee T\bigvee_{i=k+1}^{c,\alpha} \beta_i\right)$ is also equivalent to $T\beta_{k_0}$. The equivalence in the first item:

$$T \bigvee_{i=k}^{c,\alpha} \beta_i \equiv T\beta_{k_0}$$

follows for $k = k_0$.

We have shown the first equivalence for $k = k_0$. Let us now prove the induction step assuming that the equivalence holds for $k+1$. Suppose that $T\bigvee_{i=k}^{c,\alpha} \beta_i$. Then the formula in the second bracket is true as well. By minimality of k, it is not the case that $T\alpha_k$, and thus it must be the case that

$$T \bigvee_{i=k+1}^{c,\alpha} \beta_i.$$

By induction hypothesis, this implies $T\beta_{k_0}$.

Conversely, if $T\beta_{k_0}$ holds, then by induction hypothesis

$$T \bigvee_{i=k+1}^{c,\alpha} \beta_i.$$

From which it follows that the formula in the second bracket in the definition of $\bigvee_{i=k}^{c,\alpha} \beta_i$ is true. By minimality of k_0 and the fact that for all i, either α_i or $\neg\alpha_i$ is true, we have:

$$T\neg\alpha_k,$$

from which it follows that the formula in the first bracket in the definition of $T\bigvee_{i=k+1}^{c,\alpha} \beta_i$ is true. Therefore:

$$T \bigvee_{i=k}^{c,\alpha} \beta_i.$$

The second item is proved in a similar fashion. □

3.3. Models of CT^-

In this section, we will investigate what conditions are imposed on models of PA, if we assume that they carry a truth predicate satisfying the axioms of CT^-. The general outcome will be that CT^- is semantically a very strong theory. Indeed, the main result of this section is that every model of CT^- admits an expansion to a model of UTB.

3.3.1. Lachlan's Theorem

The first result showing that CT^- imposes nontrivial semantic conditions on models of PA was the following theorem of Lachlan proved in [Lachlan, 1981]:

Theorem 98 (Lachlan)**.** *Suppose that (M, T) is a model of CT^- and M is nonstandard. Then M is recursively saturated.*

The proof of Lachlan was very surprising and rather tricky. In this section, we will present a slightly modified version of his proof, since we think that it nicely illuminates the structure of our argument that every model of CT^- admits an expansion to a model of UTB.

Actually, our theorem has been preceded by a related result by Smith from [Smith, 1989].

Theorem 99 (Smith)**.** *Suppose that (M, T) is a model of CT^-. Then there exists $T' \subset M$ such that $(M, T') \models \mathrm{UTB}_0$, where UTB_0 is UTB^- enriched with the induction scheme for the Δ_0-formulae containing the truth predicate.*

In particular, such a model M carries an undefinable class, i.e. an undefinable subset whose all initial segments are coded. This was the original formulation of Smith's theorem. Consequently, not every recursively saturated model of PA expands to a model of CT^-. Basically, what we do, is to improve the methods employed by Smith so that we can get a fully inductive truth predicate rather than simply Δ_0-inductive one. The main obstacle is that although we can finitely axiomatise over PA what does it mean for a truth predicate to satisfy Δ_0-induction, we cannot finitely axiomatise the requirement that it satisfies the full induction scheme.

Now, we will prove Lachlan's theorem. Although the idea of the proof is not original, nor is the present argument really simplified compared to the original one, we hope that our presentation is more perspicuous and that it will make the proof seem more natural and better motivated.

We will show that every model which admits an expansion to a model of CT^- is recursively saturated. Let us begin with an auxiliary notion of a rank of formulae.

Definition 100. Let (M, T) be any model of CT^-. Let $p(x) = \{\phi_i(x) \mid i \in \omega\}$ be any countable type with parametres and let $\psi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}(M)$. We say that the p-rank of ψ is $n \in \omega$, if n is the least natural number such that

$$(M, T) \not\models \forall x \left(\bigwedge_{i=0}^{n} T\psi(\underline{x}) \to \phi_i(x) \right) \wedge \exists y \, T\psi(\underline{y}).$$

We say that the rank of a formula ψ is ∞, if it is not equal to any $n \in \omega$. Note that if $\psi(x)$ does not imply $\phi_i(x)$ for any i, then the rank is 0. We denote the p-rank with $\mathrm{rk}_p(\phi)$. We say that one formula has greater rank than the other meaning the obvious order on the structure $\omega \cup \{\infty\}$.

The notion defined above is very natural in our context. It is clearly enough to show for every recursive type p with parametres that there exists a formula ψ with $\mathrm{rk}_p(\psi) = \infty$. The next lemma encapsulates the technical core of the proof of Lachlan's Theorem.

Lemma 101. *Let (M, T) be a nonstandard model of CT^- and let $d \in M$ be a nonstandard element. Suppose that $r : M \cap [0, d] \to \omega \cup \{\infty\}$ is a function such that for any $a < d$ either $r(a) = \infty$ or*

$$r(a) < r(a + 1).$$

Then there exists some $b \in M$ with $r(b) = \infty$.

Proof. Take an arbitrary nonstandard $b' \in M$. If r does not attain ∞ on $[0, b']$, then $r(\gamma_{b'}) > r(\gamma_{b'-1}) > r(\gamma_{b'-2}) > \ldots$ which would yield an infinite descending chain in the well-ordering $\omega \cup \{\infty\}$. □

In this section, we will use the above lemma for $r = \mathrm{rk}_p$. In the next one, it will be applied once again, with a different notion of rank. We believe that generalising the lemma from $\omega \cup \{\infty\}$ to other well-founded relations may be possibly fruitful for the study of models of CT^-.

Now, the only thing we have to do is to show that in every nonstandard model $(M, T) \models \mathrm{CT}^-$ and for any recursive type p, one can indeed find a family of formulae γ_i such that the function $r(a) = \mathrm{rk}_p(\gamma_a)$ satisfies the assumptions of Lemma 101. With disjunctions with stopping condition at hand, this is surprisingly easy.

Lemma 102 (Rank Lemma for rk_p)**.** *Let $(M, T) \models \mathrm{CT}^-$ be a nonstandard model and let $p = \{\phi_i \mid i \in \omega\}$ be a recursive type where ϕ_i-s are arithmetical formulae with parametres. Then there exists a coded sequence $(\gamma_i)_{i=0}^{c}$ such that for an arbitrary $a < c$, if $\mathrm{rk}_p(\gamma_a) \neq \infty$, then*

$$\mathrm{rk}_p(\gamma_a) < \mathrm{rk}_p(\gamma_{a+1}).$$

Proof. We define the formulae γ_i by induction on i. We set $\gamma_0(x) = (x = x)$. The choice is, however, completely arbitrary. Suppose that we have already defined the formula γ_i. Then let

$$\alpha_j^{\gamma_i} = \exists y \big(\gamma_i(y) \wedge \neg\phi_j(y)\big) \vee \neg\exists y\ \gamma_i(y).$$

Although, γ_i has been actually *used* in the definition of $\alpha_j^{\gamma_i}$, let us denote it as $\alpha_j(\gamma_i)$ for typographical reasons. The sequence $(\alpha_j(\gamma_i))_{j=0}^{\infty}$ will be denoted with $\alpha(\gamma_i)$. Let for all $i \geq 0$,

$$\beta_i(x) = \bigwedge_{j=0}^{i+1} \phi_j(x).$$

Let finally

$$\gamma_{i+1}(x) = \bigvee_{i=0}^{c,\alpha(\gamma_i)} \beta_i(x).$$

Since the definition of the formulae γ_i is primitive recursive, it can be formalised, so that we get a coded sequence $(\gamma_i)_{i=0}^{d}$ of nonstandard length. We will show that it satisfies the claim of the lemma.

Choose any $a < d$. Suppose that $\mathrm{rk}_p(\gamma_a) = n \in \omega$. Then by definition of α_n, $(M,T) \models T\alpha_n$ and it is the least such n. Hence by Lemma 97, $\gamma_{a+1}(x)$ is equivalent to β_{n+1} which obviously has rank at least $n+1$. □

As a direct corollary of the Lemmata 101 and 102, we obtain Lachlan's Theorem 98.

3.3.2. CT$^-$ and UTB

In this subsection we shall prove a common strengthening of Lachlan's Theorem 98 and Smith's Theorem 99. Our argument will actually very closely follow that of Smith although we hope the reader will find our presentation in terms of disjunctions with stopping condition somewhat easier to follow. The proof is parallel to the one presented in the previous subsection. We basically only have to introduce a different notion of the rank. This allows us to make the presentation of our theorem much cleaner than in our original paper [Łełyk and Wcisło, 2017a].

Theorem 103. *Let $(M,T) \models \mathrm{CT}^-$. Then there exists $T' \subset M$ such that $(M,T') \models \mathrm{UTB}$.*

As we have mentioned above, we will first introduce a suitable notion of a rank. We shall need a piece of notation. Let $(\phi_i)_{i=0}^{\infty}$ be any primitive recursive enumeration of all arithmetical formulae (not necessarily of one variable) and let $(\mathrm{ind}_i)_{i=0}^{\infty}$ be any primitive recursive enumeration of the instances of the induction scheme in the language $\mathscr{L}_{\mathrm{PA}P}$, i.e. the language of arithmetic enriched with one fresh predicate $P(v)$.

Definition 104. Let (M,T) be any model of CT$^-$. Let $\psi \in \mathrm{Form}_{\mathrm{PA}}^{\leq 1}(M)$. We say that the UTB-rank of ψ is n, if it is the least natural number such that

$$(M,T) \not\models T\Big(\mathrm{ind}_n[\psi] \wedge \forall x_1,\dots,x_k\Big(\psi\big(\phi_n(\underline{x_1},\dots,\underline{x_k})\big) \equiv \phi_n(x_1,\dots,x_k)\Big)\Big).$$

If such a number n does not exist, we will say that the rank of ψ is ∞. We endow the set $\omega \cup \{\infty\}$ with the obvious order. We will denote the UTB-rank of ψ simply with $\mathrm{rk}(\psi)$.

Just as in the previous subsection, the rank function measures how close our formulae get to satisfying the claim of the theorem. In our case, if $\mathrm{rk}(\psi) = \infty$, then Theorem 103 is satisfied with $T' = T * \psi$. Therefore, by Lemma 101, the only thing that we have to check is that there indeed exists a family of formulae for which the newly introduced rank is locally growing. To this end we will again use disjunctions with stopping conditions.

Lemma 105. *Let (M,T) be any nonstandard model of* CT^-*. Then there exists a coded sequence $(\gamma_i)_{i=0}^d$ of nonstandard length such that for every $a \leq d$, $\gamma_a \in \mathrm{Form}_{\mathrm{PA}}^{\leq 1}(M)$ and if $a < d$, then either* $\mathrm{rk}(\gamma_a) = \infty$ *or*

$$\mathrm{rk}(\gamma_a) < \mathrm{rk}(\gamma_{a+1}).$$

Proof. Let (M,T) be a nonstandard model of CT^-. We will construct the formulae γ_i by induction using disjunctions with stopping condition. For any arithmetical formula ψ and a natural number n, let

$$\alpha_n^\psi = \neg\Big(\mathrm{ind}_n[\psi] \wedge \forall x_1, \ldots, x_k \Big(\psi\big(\phi_n(\underline{x_1}, \ldots, \underline{x_k})\big) \equiv \phi_n(x_1, \ldots, x_k)\Big)\Big).$$

Basically, α_n^ψ expresses that the rank of ψ is no greater than n. The formula ψ is actually used in α_i^ψ, so it cannot be treated as a variable. For typographical reasons, let us however still write it as $\alpha_i(\psi)$ and denote the sequence $(\alpha_i(\psi))_{i=0}^\infty$ with $\alpha(\psi)$. Now, let Θ_n be a formula that works like a naïve truth predicate for the formulae $\phi_1, \ldots, \phi_n$, i.e.,

$$\Theta_n(x) = \bigvee_{i=0}^{n} \exists t_1, \ldots, t_{k(i)} \in \mathrm{ClTerm}_{\mathrm{PA}}\Big(x = \phi_i(t_1, \ldots, t_{k(i)}) \wedge \phi_i({t_1}^\circ, \ldots, {t_{k(i)}}^\circ\Big),$$

where $k(i)$ is the number of free variables in ϕ_i. Note that Θ_n is defined exactly so that

$$\mathrm{rk}(\Theta_n) > n$$

(since Θ_n for $n \in \omega$ is a standard arithmetical formula, it automatically satisfies all the instances of the induction scheme). Finally, we let

$$\begin{aligned} \gamma_0(x) &= (x = x) \\ \gamma_{n+1}(x) &= \bigvee_{i=0}^{d,\alpha(\gamma_n)} \Theta_i(x). \end{aligned}$$

In effect, the formula γ_{n+1} is defined as follows: find what is the rank of γ_n and let γ_{n+1} be the canonical formula of a strictly higher rank. Since the definition of γ_i is primitive recursive, we can extend it to form a sequence $(\gamma_i)_{i=0}^d$ for some nonstandard d. Now, we shall prove that this sequence satisfies the claim of the lemma.

Obviously, all γ_a belong to $\mathrm{Form}_{\mathrm{PA}}^{\leq 1}$. Fix any $a < d$ and suppose that $\mathrm{rk}(\gamma_a) = n$. This means that n is the least natural number such that

$$(M,T) \not\models T\Big(\mathrm{ind}_n[\gamma_a] \wedge \forall x_1, \ldots, x_k \Big(\gamma_a\big(\phi_n(\underline{x_1}, \ldots, \underline{x_k})\big) \equiv \phi_n(x_1, \ldots, x_k)\Big)\Big).$$

Then by definition, n is also the least number such that $T\alpha_n(\gamma_a)$ holds. By Lemma 97 on disjunctions with stopping conditions, the following holds:

$$(M,T) \models \forall x \, \big(T * \gamma_{a+1}(x) \equiv T * \Theta_n(x) \equiv \Theta_n(x)\big).$$

By Lemma 50 on generalised commutativity, the rank of γ_{a+1}, i.e., the least number n such that

$$(M,T)\not\models T\Big(\mathrm{ind}_n[\gamma_{a+1}]\wedge\forall x_1,\ldots,x_k\Big(\gamma_{a+1}\big(\phi_n(\underline{x_1},\ldots,\underline{x_k})\big)\equiv\phi_n(x_1,\ldots,x_k)\Big)\Big),$$

is equal to the least n such that

$$(M,T)\not\models \mathrm{ind}_n[T*\gamma_{a+1}]\wedge\forall x_1,\ldots,x_k\Big(T*\gamma_{a+1}\big(\phi_n(\underline{x_1},\ldots,\underline{x_k})\big)\equiv\phi_n(x_1,\ldots,x_k)\Big).$$

Which by the above considerations is also the least number n such that

$$(M,T)\not\models \mathrm{ind}_n[\Theta_n]\wedge\forall x_1,\ldots,x_k\Big(\Theta_n\big(\phi_n(\underline{x_1},\ldots,\underline{x_k})\big)\equiv\phi_n(x_1,\ldots,x_k)\Big).$$

Which by definition equals to $\mathrm{rk}(\Theta_n)$. And, as we have already observed before,

$$\mathrm{rk}(\Theta_n)>n=\mathrm{rk}(\gamma_a).$$

□

Proof of Theorem 103. Let $(M,T)\models \mathrm{CT}^-$. If M is standard, then every predicate in M satisfies the full induction scheme, so in particular $(M,T)\models \mathrm{UTB}$. Suppose that M is nonstandard. Then, exactly as in the case of Lachlan's theorem, Lemmata 101 and 105 imply that for some $a\leq d$ the formula γ_a has rank equal to ∞. Therefore, again by a simple use of Generalised Commutativity Lemma 50 and Internal–External Lemma 53, we conclude that $(M,T')\models \mathrm{UTB}$ for $T':=T*\gamma_a$. □

Chapter 4

Model-theoretic strength II: positive truth

This chapter will be devoted to investigating the models of compositional theories of truth whose axioms are not modelled after classical logic, but rather some form of partial logic. Theories with such compositional axioms have first been investigated in the context of self-referential truth predicates.[1] If the self-referentiality is dropped and the predicates only refer to arithmetical sentences, then the resulting theories tend to be in general weaker than their classical counterparts, so they form a nice source of examples of weak theories of truth which display some features typical of strong theories.

4.1. Positive compositional truth with total internal induction

The most basic result, which shows that positive compositional truth theories tend to be weaker than their classical counterparts states that PT^- is semantically not stronger than PA itself, in dramatic contrast to the case of CT^-.

Proposition 106. PT^- *is semantically conservative over* PA. *Thus* $\mathrm{PT}^- <_{\mathrm{mod}} \mathrm{CT}^-$.

Proof. Let M be an arbitrary model of PA and let us define the following operator Γ :

[1] For a general introduction to truth theories with self-referential truth predicates, see Halbach's monograph [Halbach, 2011], Part III.

$\mathcal{P}(M) \to \mathcal{P}(M)$.

$$x \in \Gamma(X) \equiv \begin{cases} \exists s,t \in \mathrm{ClTerm}_{\mathrm{PA}}\ x = (s = t) & \wedge & s^\circ = t^\circ \\ \vee \exists s,t \in \mathrm{ClTerm}_{\mathrm{PA}}\ x = (s \neq t) & \wedge & s^\circ \neq t^\circ \\ \vee \exists \phi \in \mathrm{Sent}_{\mathrm{PA}}\ x = \neg\neg\phi & \wedge & \phi \in X \\ \vee \exists \phi,\psi \in \mathrm{Sent}_{\mathrm{PA}}\ x = \phi \wedge \psi & \wedge & \phi \in X \wedge \psi \in X \\ \vee \exists \phi,\psi \in \mathrm{Sent}_{\mathrm{PA}}\ x = \neg(\phi \wedge \psi) & \wedge & ((\neg\phi) \in X \vee (\neg\psi) \in X) \\ \vee \exists \phi,\psi \in \mathrm{Sent}_{\mathrm{PA}}\ x = \phi \vee \psi & \wedge & (\phi \in X \vee \psi \in X) \\ \vee \exists \phi,\psi \in \mathrm{Sent}_{\mathrm{PA}}\ x = \neg(\phi \vee \psi) & \wedge & ((\neg\phi) \in X \wedge (\neg\psi) \in X) \\ \vee \exists \phi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}, v\ x = \exists v\ \phi(v) & \wedge & \text{for some } x \in M\ \phi(\underline{x}) \in X \\ \vee \exists \phi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}, v\ x = \neg\exists v\ \phi(v) & \wedge & \text{for all } x \in M\ (\neg\phi(\underline{x})) \in X \\ \vee \exists \phi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}, v\ x = \forall v\ \phi(v) & \wedge & \text{for all } x \in M\ \phi(\underline{x}) \in X \\ \vee \exists \phi \in \mathrm{Form}^{\leq 1}_{\mathrm{PA}}, v\ x = \neg\forall v\ \phi(v) & \wedge & \text{for some } x \in M\ (\neg\phi(\underline{x})) \in X. \end{cases}$$

As one can read off the definition, the operator Γ is monotone with respect to the inclusion order $\subseteq$. Therefore we can iterate it through the transfinite numbers by setting:

$$\begin{aligned} \Gamma^0(X) &= X \\ \Gamma^{\alpha+1}(X) &= \Gamma(\Gamma^\alpha(X)) \\ \Gamma^\gamma(X) &= \bigcup_{\beta<\gamma} \Gamma^\beta(X), \text{ for limit numbers } \gamma. \end{aligned}$$

Since Γ is monotone, for every α,

$$\Gamma^\alpha(\emptyset) \subseteq \Gamma^{\alpha+1}(\emptyset),$$

thus, by cardinality reasons, there is some ordinal α such that $T := \Gamma^{\alpha+1}(\emptyset) = \Gamma^\alpha(\emptyset)$. One can check that since T is a fixpoint of the operator Γ, the model (M, T) satisfies PT^-. □

There was a hope that the theory PT^- may be strengthened a bit, so that it still is model-theoretically conservative, but gains a superexponential speed-up over PA. Let us define this important notion which we have already mentioned in the remarks following the proof of Theorem 60.

Definition 107. Let $\mathrm{Th} \subseteq \mathrm{Th}'$ be two theories and let $\mathscr{L}$ be the language of Th. We say that Th' has a **superexponential speed-up** over Th, if there exists a sequence $(\phi_i)_{i=0}^\infty$ of formulas in the language $\mathscr{L}$ such that for any function f which may be obtained by composition from polynomials and the exponential function x^y, and for sufficiently large i, the following inequality holds:

$$|\phi_i|_{\mathrm{Th}} \geq f\left(|\phi_i|_{\mathrm{Th}'}\right),$$

where $|\phi|_S$ means a number of symbols of the smallest proof of ϕ in the the theory S.

Now, we shall define a theory which indeed may be shown to have a super-exponential speed-up over PA. It was also hoped that this theory will turn out to be model-theoretically conservative over PA.

Definition 108. By **internal induction for total formulae** we mean the following axiom:

$$\forall \psi(v) \in \mathrm{Form}_{\mathrm{PA}} \Big[\mathrm{tot}(\psi) \longrightarrow \Big(\forall x \Big(T{*}\psi(x) \to T{*}\psi(Sx) \Big) \longrightarrow \Big(T{*}\psi(0) \to \forall x\, T{*}\psi(x) \Big) \Big) \Big], \tag{INTtot}$$

where $\mathrm{tot}(\phi)$ means that ϕ is **total**, i.e., the following holds:

$$\forall x\ (T\phi(\underline{x}) \vee T\neg\phi(\underline{x})).$$

Totality of formulae is an extremely natural notion in the context of truth theories based on partial logics. A formula is total if it is either true or false of every object. In a sense, totality of a formula is tantamount to its defining a well-behaved property for which there are no undecided cases. In a way, INTtot requires that those well-behaved properties satisfy induction. Having written that, we admit that the axiom INTtot is rather technical.

It was hoped and even claimed in the literature[2] that PT$^-$+ INTtot is semantically conservative and has a speed-up over PA. In other words, this theory was claimed to be at the same time *useful* as a tool and innocent, since it does not put any restrictions on models of PA. The speed-up may be indeed proved using techniques developed by Friedman, Solovay, Vopěnka, and Pudlák (see [Pudlák, 1998] and [Pudlák, 1986], Corollary 4.1). For this specific theory, it has been demonstrated by Fischer in [Fischer, 2014], Theorem 9.

Theorem 109 (Fischer)**.** PT$^-$+ *INTtot has superexponential speed-up over* PA.

Unfortunately, it turned out that the proof of semantic conservativity of PT$^-$+ INTtot presented in [Fischer, 2009] was flawed. The gap has been spotted independently by Cezary Cieśliński and Carlo Nicolai together with Albert Visser. Moreover, it turned out that the gap was essential, since the conservativity claim fails.[3]

Theorem 110 (Cieśliński–Łełyk–W)**.** *Let* $M \models$ PA *be a nonstandard model such that there exists an expansion* $(M, T) \models$ PT$^-$+ *INTtot. Then* M *is short recursively saturated. In particular,* PT$^-$+ *INTtot is not semantically conservative over* PA.

Proof. Fix any model $(M, T) \models$ PT$^-$+ INTtot and any recursive type $p = \{\phi_i(x) \mid i \in \omega\}$ over M such that for some a and an arbitrary $n \in \omega$,

$$(M, T) \models \exists x < a \bigwedge_{i=0}^{n} \phi_i(x).$$

We will show that the type p is realised below a. Suppose that this type is omitted and let us define:

$$\alpha_n(y) = \neg\big(y < \underline{a} \wedge \phi_0(y) \wedge \phi_1(y) \wedge \ldots \wedge \phi_n(y)\big).$$

[2] In [Fischer, 2009], results under the heading "Some semantic conservativity results" on p. 804.

[3] See [Cieśliński et al., 2017], Theorem 4.

Note that this is similar to what we have defined in the proof of Lachlan's Theorem, or more precisely, in the proof of Lemma 102. Let

$$\beta_n(x) = x < \underline{a} \wedge \phi_0(x) \wedge \phi_1(x) \wedge \ldots \wedge \phi_{n+1}(x).$$

Now, let us fix any nonstandard c and let

$$\psi(x,y) = \bigvee_{j=0}^{c,\alpha(y)} \beta(x).$$

Intuitively, the formula $\psi(x,y)$ expresses the fact that x satisfies a bigger portion of the type p than y does. Finally, let

$$\eta(z) = \exists x \forall y < z \; \psi(x,y).$$

In other words, $\eta(z)$ expresses that there is an element that realises a larger portion of the type p than any of the elements below z.

We claim that the formula $\eta(z)$ is total. It is enough to show that the formula $\psi(x,y)$ is total (i.e., for arbitrary x,y, exactly one of $T\psi(\underline{x},\underline{y})$, $T\neg\psi(\underline{x},\underline{y})$ holds). Obviously for any $j,k \in \omega$, the formulae α_j, β_k are total, since they are standard (see Proposition 51). Since the type p is omitted, for every y, there exists some $k \in \omega$ such that $T\alpha_k(\underline{y})$. If we denote that k by $k(y)$, then by Lemma 97, we obtain the following equivalence:

$$\forall x \left(T\psi(\underline{x},\underline{y}) \equiv \beta_{k(y)}(x) \right),$$

which implies that ψ is total.

Now we claim that for any $z \in M$,

$$(M,T) \models T\eta(\underline{z}) \rightarrow T\eta(\underline{z+1}).$$

Fix any $z \in M$. Suppose that $(M,T) \models T\eta(\underline{z})$. This means that for some $x \in M$,

$$(M,T) \models \forall y < z \; T\psi(\underline{x},\underline{y}).$$

Consider $k(z)$. Since p is a type, there exists some x' such that $(M,T) \models \beta_{k(z)}(x')$, i.e., x' satisfies a greater portion of the type p than z does.

Now, let x'' equal to x if $\beta_{k(z)}(x)$ holds and let it equal to x' otherwise. In other words, we define x'' to be either x or x', whichever satisfies "more" formulae from the type p. Then

$$(M,T) \models \forall y < z+1 \; T\psi(\underline{x''},\underline{y}),$$

which proves the claim.

One can readily check that $T\eta(\underline{0})$ holds vacuously. Therefore by internal induction for the total formula η, we may conclude that $(M,T) \models \forall z \; T\eta(\underline{z})$. In particular,

$$(M,T) \models T\eta(\underline{a}).$$

This means that there exists $b \in M$ such that

$$(M, T) \models \forall y < a \, T\psi(\underline{b}, \underline{y}).$$

Now, for arbitarily large $n \in \omega$, there exists $y < a$ such that $\phi_i(y)$ holds for all $i \leq n$ and does not hold for $i = n + 1$. Therefore, $(M, T) \models T\alpha_{n+1}(\underline{y})$, so by the choice of b we conclude that $(M, T) \models T\beta_{n+1}(\underline{b})$. Therefore, we conclude that for all $k \in \omega$,

$$(M, T) \models T\beta_k(\underline{b}).$$

This means that b realises p in the model M. Since this holds for an arbitrary p, M is short recursively saturated. □

Note that in the above proof, we actually used the fact that the type p is finitely realised below a. We concluded that there exists an element which realises all formulae from p from the fact that there exists an element which realises more formulae than any other element below a. If the type were not finitely satisfied below a, this step would break down. It may seem that this assumption is not really essential and some easy fix would let us lift this argument to more general setting and show that any model M of PA which expands to a model of PT$^-$+ INTtot is recursively saturated but, as of this moment, we have found no such fix.

For all what we know, such an easy fix may in fact exist. We know however that we will not be able to obtain more than recursive saturation thanks to a result of Mateusz Łełyk announced in [Cieśliński et al., 2017], Theorem 10.

Theorem 111 (Łełyk)**.** *Suppose that $M \models$ PA is recursively saturated. Then M admits an expansion to a model $(M, T) \models$ PT$^-$+ INTtot.*

Thus we know that the class of nonstandard models of PA which admit an expansion to a model of PT$^-$+ INTtot is contained between the class of short recursively saturated models and recursively saturated models. Currently, it is not clear to us whether any of the two kinds of saturations and expandability to the model of the discussed theory will turn out to coincide.

It is worth stressing that Theorem 111 holds independently of the cardinality of the model M in question which makes PT$^-$ + INTtot a rather weak theory of truth from the point of view of the semantic strength.

4.2. Positive compositional truth with unrestricted internal induction

In the previous section, we have shown that the theory PT$^-$+ INTtot is not semantically conservative over PA, i.e., it is nontrivial from the model-theoretic point of view. The reader might be concerned that the internal induction axiom has been restricted in the investigated theory in a poorly justified way. We may dismiss the objection, simply by reminding that since we have shown that the theory in question is *strong*, its counterpart with the unrestricted induction is strong as well. Nevertheless, we admit that

the primary reason of our interest in PT$^-$+ INTtot was that it was already discussed in the literature on truth.

In this subsection, we will investigate into the semantic strength of PT$^-$+ INT, i.e. a theory with the positive compositional axioms for the truth predicate and the unrestricted internal induction axiom. A question of particular interest is whether lifting the restriction on formulae for which internal induction is valid will in any way affect the semantic strength of the obtained theory.

It turns out that PT$^-$+ INT is in fact stronger than PT$^-$+ INTtot. Moreover, it turns out to be at least as strong as UTB. This result has been obtained in collaboration with Mateusz Łełyk.

Theorem 112 (Łełyk–W)**.** *Suppose that $M \models$ PA expands to a model $(M, T) \models$ PT$^-$+ INT. Then it expands to a model $(M, T') \models$ UTB.*

Fix a model $(M, T) \models$ PT$^-$+INT. Using some parametres from M, we will find a formula which will satisfy the axioms of UTB, i.e. the uniform Tarski's biconditionals of UTB$^-$ along with the full induction scheme.

Like in Subsection 3.3.2, let $(\phi_i)_{i=0}^{\infty}$ be a primitive recursive enumeration of arithmetical formulae, possibly with many variables. Without loss of generality, we may assume that M is nonstandard. Let us fix an arbitrary nonstandard $c \in M$. Let us define ternary formulae $\alpha_i(x, y, z)$ as follows:

$$\alpha_i(x, \bar{t}, b) = \underline{i} \geq b \vee \Big(\bar{t} \in \mathrm{ClTermSeq}_{\mathrm{PA}} \wedge x = \phi_i(\bar{t})\Big).$$

The idea behind formula $\alpha_i(x, y, b)$ is almost trivial. Its second part expresses that x results by substituting some closed terms in the i-th formula in the fixed enumeration. The first part is simply an additional constraint of which formulae are considered at all (which will play a crucial role once we place α_i's in a disjunction with stopping conditions). Note that $\bar{t}$ is a suggestive name of a variable intended to denote a sequence, not for a sequence of variables.

Let us define binary formulae $\beta_i(x, y)$ as follows:

$$\beta_i(\bar{t}, b) = \underline{i} < b \wedge \mathrm{lh}(\bar{t}) \geq \underline{l} \wedge \phi_i\big(t_1{}^\circ, \ldots, t_l{}^\circ\big),$$

where l is the number of free variables in the formula ϕ_i. The second part of the formula β_i simply says that if we substitute terms from the sequence $\bar{t}$ in the i-th formula of our fixed enumeration, then the resulting sentence is true. The first part is really technical (yet simple) and the reader is advised to wait and see how it will actually be used.

Let us define an indexed family of formulae $\psi_c(x, y)$ as follows:

$$\psi_c(\phi, b) = \exists \bar{t} \in \mathrm{TermSeq}_{\mathrm{PA}} \bigvee_{i=0}^{c, \alpha_i(\phi, \bar{t}, b)} \beta_i(\bar{t}, b).$$

Now, the formulae ψ_c as defined above already are enough to define a truth predicate satisfying UTB$^-$.

Lemma 113. *Let $(M,T) \models \mathrm{PT}^-$, let $b, c \in M$ be arbitrary two nonstandard elements. Then for an arbitrary standard arithmetical formula ϕ,*

$$(M,T) \models \forall t_1, \ldots, t_n \in \mathrm{ClTerm}_{\mathrm{PA}} \Big(T * \psi_c\big(\phi(t_1, \ldots, t_n), b\big) \equiv \phi\big({t_1}^\circ, \ldots, {t_n}^\circ\big) \Big).$$

Proof. Fix any $(M,T) \models \mathrm{PT}^-$, a standard arithmetical formula ϕ, and $t_1, \ldots, t_n \in \mathrm{ClTerm}_{\mathrm{PA}}$. Since ϕ is standard, there exists the least $i \in \omega$ such that for some $\overline{s} \in \mathrm{TermSeq}_{\mathrm{PA}}$

$$\phi(t_1, \ldots, t_n) = \phi_i(\overline{s}).$$

Then $(M,T) \models T * \alpha_i(\phi(t_1, \ldots, t_n), \overline{s}, b)$ holds (we have assumed that b is nonstandard, so $b > i$). Hence by Lemma 97, we may conclude that

$$(M,T) \models T * \psi_c(\phi(t_1, \ldots, t_n), b) \equiv T\big(\phi_i(\underline{{s_1}^\circ}, \ldots, \underline{{s_k}^\circ})\big) \equiv \phi({t_1}^\circ, \ldots, {t_n}^\circ),$$

where $k \in \omega$ is the length of the sequence $\overline{s}$. The last equivalence holds, since $\phi_i(s_1, \ldots, s_k)$ and $\phi(t_1, \ldots, t_n)$ as codes of sentences are literally the same object. This concludes the proof. □

The next lemma explains why we need a parameter b in the definition of ψ.

Lemma 114. *Let $(M,T) \models \mathrm{PT}^- + \mathit{INT}$, $c \in M$. Take any $b \in M$ such that*

$$(M,T) \models \forall x \Big(\Big(T\psi_c(\underline{x}, \underline{b}) \vee T\neg\psi_c(\underline{x}, \underline{b}) \Big) \wedge \neg \Big(T\psi_c(\underline{x}, \underline{b}) \wedge T\neg\psi_c(\underline{x}, \underline{b}) \Big) \Big).$$

Then the predicate $T'(x) = T * \psi_c(x, b)$ satisfies the full induction scheme.

Proof. For an arbitrary $\phi \in \mathscr{L}_{\mathrm{PA}P}$, a formula with a fresh unary predicate P, in any model $(M,T) \models \mathrm{PT}^- + \mathrm{INT}$ the following holds:

$$\forall x \, \Big(T(\phi[\psi_c](\underline{x}, \underline{b})) \to T(\phi[\psi_c](\underline{Sx}, \underline{b})) \Big) \longrightarrow \Big(T(\phi[\psi_c](\underline{0}, \underline{b})) \to \forall x \, T(\phi[\psi_c](\underline{x}, \underline{b})) \Big).$$

Then the claim follows by Internal–External Lemma 53, once we show that T is compositional across (Φ, ξ), where $\xi(x) = \psi(x, \underline{b})$ (with b fixed) and Φ is the above instance of the induction scheme. However, $\xi(x)$ is total and consistent and $\phi \in \mathscr{L}_{\mathrm{PA}P}$ is standard and we can then check by simple induction on complexity of formulae that in such case T is compositional between ξ and $\phi[\xi]$ and, consequently, between ξ and Φ. □

Now we only have to show that in any model (M,T) we can indeed find a parameter b, for which the formula $\psi(x, b)$ will be total and consistent. This is easily achieved by another application of the internal induction axiom.

Lemma 115. *Let $(M,T) \models \mathrm{PT}^- + \mathit{INT}$. Then for an arbitrary c there exists a nonstandard b such that*

$$(M,T) \models \forall x \Big(T\psi_c(\underline{x}, \underline{b}) \vee T\neg\psi_c(\underline{x}, \underline{b}) \Big) \wedge \neg \Big(T\psi_c(\underline{x}, \underline{b}) \wedge T\neg\psi_c(\underline{x}, \underline{b}) \Big).$$

Proof. Let us fix an arbitrary nonstandard c and for the rest of the proof let us denote $\psi = \psi_c$. Note that for any standard b and any $\phi \notin \{\phi_i(\bar{t}) \mid i = 1, \ldots, b, \bar{t} \in \mathrm{ClTermSeq}_{\mathrm{PA}}\}$,

$$(M, T) \models T\neg\psi(\underline{\phi}, \underline{b}).$$

Namely, the first k such that $(M, T) \models \alpha_k(\underline{\phi}, \underline{\bar{t}}, \underline{b})$ is b. By definition, $\beta_i(x, b)$ is never satisfied for $b \geq i$. In particular, $\beta_b(x, b)$ is false for all x, hence the conclusion.

If, in turn, ϕ can be presented in the form $\phi_i(t_1, \ldots, t_n)$ for some $i = 1, \ldots, b$ and $\langle t_1, \ldots, t_n\rangle \in \mathrm{ClTermSeq}_{\mathrm{PA}}$, then

$$(M, T) \models \forall t_1, \ldots, t_n \in \mathrm{ClTerm}_{\mathrm{PA}}\Big(T * \psi\big(\phi(t_1, \ldots, t_n), b\big) \equiv \phi\big({t_1}^\circ, \ldots, {t_n}^\circ\big)\Big).$$

This may be shown exactly like Lemma 113. in particular, for every ϕ of the form $\phi_i(t_1, \ldots, t_n)$ for some $t_1, \ldots, t_n \in \mathrm{ClTermSeq}_{\mathrm{PA}}$,

$$(M, T) \models \Big(T\psi(\underline{\phi}, \underline{b}) \vee T\neg\psi(\underline{\phi}, \underline{b})\Big) \wedge \neg\Big(T\psi(\underline{\phi}, \underline{b}) \wedge T\neg\psi(\underline{\phi}, \underline{b})\Big).$$

Therefore for an arbitrary standard $b \in \omega$, the following holds:

$$(M, T) \models \forall x \Big(T\psi(\underline{x}, \underline{b}) \vee T\neg\psi(\underline{x}, \underline{b})\Big) \wedge \neg\Big(T\psi(\underline{x}, \underline{b}) \wedge T\neg\psi(\underline{x}, \underline{b})\Big).$$

Now, consider the following formula:

$$\eta_1(b) = \forall x \Big(\psi(\underline{x}, \underline{b}) \vee \neg\psi(\underline{x}, \underline{b})\Big).$$

We have just shown that $(M, T) \models T\forall y\Big(y < \underline{b} \to \eta_1(y)\Big)$ for all standard b. By internal induction, this implies that there exists a nonstandard b' such that

$$(M, T) \models T\forall y\Big(y < \underline{b'} \to \eta_1(y)\Big).$$

By compositional axioms for PT^-, this implies that

$$(M, T) \models \forall y < b'\Big(T\psi(\underline{x}, \underline{b}) \vee T\neg\psi(\underline{x}, \underline{b})\Big).$$

Let us define another formula:

$$\eta_2(z) = \exists x \exists y\Big(y \leq b' - z \wedge \psi(x, y) \wedge \neg\psi(x, y)\Big).$$

The formula above states that the formula $\psi(x, y)$ is not consistent for some fixed y below b'. It is written using the subtraction operation, since we have to be careful to use only positive compositional clauses (or to use compositional clauses for formulae which are guaranteed to be total).

Suppose that

$$(M, T) \models T\eta_2(\underline{0}).$$

On the other hand,

$$(M,T) \models \neg T\eta_2(\underline{b'}),$$

since otherwise we would have $(M,T) \models T\psi(\underline{x},\underline{0}) \wedge T\neg\psi(\underline{x},\underline{0})$, which we have excluded at the beginning of the proof for all x and all standard b, including $b = 0$. Therefore by the internal induction axiom, there exists some z such that $b'' = b - z$ is nonstandard and

$$(M,T) \models \neg T\eta_2(\underline{z}),$$

i.e.,

$$(M,T) \models \neg T\Big(\exists x \exists y\Big(y \leq \underline{b''} \wedge \psi(x,y) \wedge \neg\psi(x,y)\Big)\Big).$$

By contraposition, we may conclude that for all $y \leq b''$ and for all x,

$$(M,T) \models \neg\Big(T\psi(\underline{x},\underline{y}) \wedge T\neg\psi(\underline{x},\underline{y})\Big).$$

We conclude that b'' is a nonstandard element satisfying the claim of the lemma, i.e.,

$$(M,T) \models \forall x\Big(T\psi(\underline{x},\underline{b''}) \vee T\neg\psi(\underline{x},\underline{b''})\Big) \wedge \neg\Big(T\psi(\underline{x},\underline{b''}) \wedge T\neg\psi(\underline{x},\underline{b''})\Big).$$

□

Now we may finish the proof of Theorem 112.

Proof of Theorem 112. Let (M,T) be a model of PT$^-$+ INT. We will show that there exists $T' \subset M$ such that $(M,T') \models$ UTB. Without loss of generality, we may assume that M is nonstandard. Fix some nonstandard $c \in M$ and use it to define $\psi_c(x,y) \in \mathrm{Form}_{\mathrm{PA}}(M)$ as above. By Lemma 115, there exists some b such that

$$(M,T) \models \forall x\Big(T\psi_c(\underline{x},\underline{b}) \vee T\neg\psi_c(\underline{x},\underline{b})\Big) \wedge \neg\Big(T\psi_c(\underline{x},\underline{b}) \wedge T\neg\psi_c(\underline{x},\underline{b})\Big).$$

By Lemma 114, the predicate $T'(x) = T * \psi_c(x,b)$ satisfies the full induction scheme. By Lemma 113, it satisfies the uniform Tarski's scheme UTB$^-$. This means that

$$(M,T') \models \text{UTB},$$

which concludes the proof. □

In the previous section, we have seen that any recursively saturated model of PA may be expanded to a model of PT$^-$+ INTtot. On the other hand, Theorem 93 states that not every recursively saturated model of PA may be expanded to a model of UTB. In effect, we obtain the following corollary:

Corollary 116. PT$^-$+ *INT is semantically stronger than* PT$^-$+ *INTtot.*

Chapter 5

Conclusions

In this chapter, we provide a summary of the results in our thesis, discuss some possible other notions of strength omitted in the main part and try to understand the relevance of our results for the philosophical debate which we have outlined in the introduction.

5.1. Summary of results

In our thesis, we have discussed two main notions of strength regarding theories of truth: proof-theoretic strength and semantic strength.

In the case of the proof-theoretic strength, we have analysed certain theories which extend a conservative theory CT^-, but which are not stronger than PA in an obvious way (as opposed to CT^-+ SPA). Here, we encounter a surprising phenomenon. It turns out that a number of natural and apparently distinct systems of axioms result in *precisely the same theory*, namely CT_0+ NORM (i.e. CT^- with induction for Δ_0 formulae containing truth predicate and the axiom stating that only sentences are true).

We discussed this surprising result in Section 2.3. Recall that CT_0 turned out to be non-conservative over PA. Thus, there exists a very robust nonconservative theory of truth which admits a number of genuinely different characterisations. This theory is equivalent to CT^-+ SPA + NORM, i.e. the compositional theory of truth with the additional axioms saying that whatever is provable in PA is true and whatever is true is a sentence.

In the case of semantic strength, we have obtained a very different, yet also elegant picture. For any two theories of truth $\mathrm{Th}_1, \mathrm{Th}_2$ whose models we managed to analyse, it turned out that actually the class of models of PA which can be expanded to a model of Th_1 and class of models of PA which can be extended to a model of Th_2 are *comparable* even when the theories themselves seem completely unrelated. The most striking example of this phenomenon was the case of UTB and CT^-. In every model of the purely compositional theory of truth, one can find an *inductive* truth predicate satisfying the scheme of uniform Tarski's biconditionals. In a sense, one can trade off compositionality for induction. Let us summarise the results on model-theoretic strength in the

following diagram:

$$\text{PA}{=}\text{PT}^- {<} \text{TB} < \text{PT}^- + \text{INTtot} < \text{UTB} \leq \text{CT}^-, \text{PT}^- + \text{INT}.$$

where $\leq, <, =$ above mean $\leq_{\text{mod}}, <_{\text{mod}}, =_{\text{mod}}$ (These notions have been introduced in Definition 88 and the remarks following it. $\text{Th}_1 \leq_{\text{mod}} \text{Th}_2$ denotes the fact that the class of models of PA which expand to a model of the theory Th_2 is contained in the class of models of PA which expand to a model of a theory Th_1. We define $<_{\text{mod}}$ and $=_{\text{mod}}$ in a similar way).

If we restrict our attention to nonstandard models, we can obtain some further information. Let us denote by **RS** and **SRS** the classes of recursively saturated and short recursively models, respectively. Let us denote by **Th** the class of nonstandard models of a theory Th. Then we can refine the above result as follows:

$$\mathbf{PA}{=}\mathbf{PT}^- {\supset} \mathbf{TB} \supset \mathbf{SRS} \supseteq \mathbf{PT}^- + \mathbf{INTtot} \supseteq \mathbf{RS} \supset \mathbf{UTB} \supseteq \mathbf{CT}^-, \mathbf{PT}^- + \mathbf{INT}.$$

Note that *prima facie* it is not at all clear that the models of the mentioned theories are comparable at all. Why should relaxing compositional axioms to axioms modelled on partial logic along with adding some induction result in a theory whose models are related to the ones of CT^-? And yet, we have thus far found no counterexample to this phenomenon among the theories of truth analysed in the literature.

Maybe it is worth stressing that in both the cases discussed we have discovered surprisingly much *connection* between various principles governing the truth predicate. We have started with a number of quite different plausible axioms for the truth predicate. What we could initially expect to obtain was somewhat chaotic and intricate picture with some theories in fact identical, some comparable in a nice way, and some really incomparable. What we have obtained instead is in both cases a very pleasant description of the relations between theories in question which looks surprisingly regular.

Let us notice that this impression may simply be due to the fact that we have analysed relatively few theories of truth and our results or methods seem rather specific to the particular examples that we have considered. Therefore, further investigation may introduce new natural systems of axioms which will not fit into the outlined pattern so neatly.

5.2. Other notions of strength

Besides the proof-theoretic strength and model-theoretic strength we discussed in this thesis, there are other possible natural notions of the strength of truth theories. In this section, let us analyse two other notions that we find particularly interesting.

5.2.1. Relative definability

Let us begin with a definition that we have already introduced in Section 2.2.

Definition 117. Let Th, Th′ be two theories of truth. We say that Th is **relatively definable** in Th′ if there exists a formula $\phi(x) \in \mathscr{L}_{\mathrm{PA}T}$ of one free variable, such that for any axiom α of Th,

$$\mathrm{Th}' \vdash \alpha[\phi/T].$$

In other words, a truth predicate satisfying the axioms of the theory Th can be defined in Th′. Arguably, this is a very fine-grained method of comparison. If Th is relatively definable in Th′, then within Th′, we can emulate the notion of truth as axiomatised by the theory Th. In particular, the following fact holds:

Fact 118. *Suppose* Th, Th′ *are two theories of truth and* Th *is relatively definable in* Th′. *Then:*

- $\mathrm{Th} \leq_{\mathscr{L}_{\mathrm{PA}}} \mathrm{Th}'$.
- $\mathrm{Th} \leq_{\mathrm{mod}} \mathrm{Th}'$.

This (easy) fact is due to Fujimoto,[1] who has also proposed relative definability as the right measure to compare the conceptual strength of distinct theories of truth. Recall that $\mathrm{Th} \leq_{\mathscr{L}_{\mathrm{PA}}} \mathrm{Th}'$ means that the arithmetical consequences of Th are contained in arithmetical consequences of Th′ and $\mathrm{Th} \leq_{\mathrm{mod}} \mathrm{Th}'$ means that any model $M \models \mathrm{PA}$ which is expandable to a model of Th′ is also expandable to a model of Th, see Definition 55 and Definition 88, respectively.

From Fact 118 and the results of this thesis, we can immediately obtain some corollaries concerning relative definability.

Corollary 119.

1. TB *does not relatively define* UTB.
2. $\mathrm{PT}^- +$ *INTtot does not relatively define* CT^-.
3. $\mathrm{PT}^- +$ *INTtot does not relatively define* $\mathrm{PT}^- +$ *INT*.
4. $\mathrm{PT}^- +$ *INTtot does not relatively define* UTB.

The first item follows by Fact 118, since every nonstandard model of UTB is recursively saturated by Proposition 92 and not every nonstandard model of TB is even short recursively saturated (Theorem 95). The second and third items follow by the fact that $\mathrm{CT}^- \geq_{\mathrm{mod}} \mathrm{UTB}$ (Theorem 103) and that $\mathrm{PT}^- + \mathrm{INT} \geq_{\mathrm{mod}} \mathrm{UTB}$ (Theorem 112), the fact that every recursively saturated model can be expanded to a model of $\mathrm{PT}^- +$ INTtot, and the fact that not every recursively saturated model can be expanded to a model of UTB (by Theorem 93). This also proves the fourth item. The first item also implies that TB^- does not relatively define UTB^-, thus answering Open Problem 2 from [Fujimoto, 2010]. It was Carlo Nicolai who first pointed out to us that this consequence follows from our model-theoretic considerations. Before that, we were not even aware of the problem.

[1] Item 2 is proved in [Fujimoto, 2010], Proposition 28 (1). Item 1 then follows by Completeness Theorem.

Note that for all theories $\mathrm{Th}_1, \mathrm{Th}_2$ such that $\mathrm{Th}_1 <_{\mathrm{mod}} \mathrm{Th}_2$, Th_1 cannot relatively interpret Th_2. Therefore, model-theoretic considerations actually yield more results on relative definability than listed in Corollary 119 (e.g., we can infer that TB does not interpret CT^-). However, the results we have not listed may be obtained with a considerably simpler argument, which also allows us to obtain some additional information not otherwise deduced from our current knowledge on models of theories of truth or simply not following from the model-theoretic argument explained above.

Proposition 120.

1. UTB *does not relatively define* CT^-.

2. UTB *does not relatively define* PT^-.

Proof. If UTB were able to define a truth predicate satisfying axioms of CT^-, this predicate would actually satisfy full induction and thus full CT. The latter theory is not conservative over PA, contrary to the former. The same argument applies to PT^-, since $\mathrm{CT} = \mathrm{PT}$ (i.e. PT^- with full induction). □

We can obtain some further negative results on relative definability based on considerations on lengths of proofs. We will prove them in the next subsection.

Proposition 121.

1. $\mathrm{PT}^- +$ *INTtot is not relatively definable in* CT^-.

2. $\mathrm{CT}^- +$ *INT is not relatively definable in* CT^-.

The latter result is an immediate corollary of the former but we list it nonetheless, since $\mathrm{CT}^- + \mathrm{INT}$ is a very natural extension of CT^-.

Let us conclude this section with some questions. It seems that the most natural proof-theoretically conservative theories of truth investigated in this thesis are UTB and CT^-. For all what we know, it may turn out that the arithmetical parts of models of these theories are exactly the same. We have already noted that UTB does not relatively define CT^- (or PT^- + INT). It seems interesting to ask whether the non-definability results reverse.

Problem 122.

1. Is UTB relatively definable in $\mathrm{CT}^- + \mathrm{INT}$?

2. Is UTB relatively definable in CT^-?

3. Is UTB relatively definable in $\mathrm{PT}^- + \mathrm{INT}$?

Note that it is by no means obvious that the answer to any of these questions is "no". By inspection of proofs of Theorems 103 and 112, one can see that we have in fact proved that, in every model of CT^- or $\mathrm{PT}^- + \mathrm{INT}$, a predicate satisfying UTB (with the same underlying arithmetical structure) is definable *with parametres*. This is quite close

to relative definability. We simply allow parameters and we allow that the formula defining the truth predicate of UTB is in general different in different models. It is not obvious whether these restrictions could or could not be dropped. In addition, note that Theorem 65 already gives an example of a nontrivial relative definability between theories of truth. Hence, lack of very direct and obvious relative definition of UTB in the theories considered above is not a good argument to the effect that there is no such definition at all. One can ask a similar question with similar caveats.

Problem 123.

1. Is TB relatively definable in CT^-+INT?
2. Is TB relatively definable in CT^-?
3. Is TB relatively definable in PT^-+ INT?
4. Is TB relatively definable in PT^-+ INTtot?

Of course, we simply ask for a weaker result than in the previous problem (except for item 4). It is not clear whether this version of the problem would be much easier to answer, even given that the solution is positive.

5.2.2. The size of proofs

Once we fix a deductive system (e.g. sequent calculus) to know what objects are valid proofs, we can introduce the notion of **size of proofs**. If ϕ is a sentence provable in a theory Th, let us denote by $|\phi|_{\mathrm{Th}}$ the minimal number of symbols occurring in a proof of ϕ in Th.[2] This leads to another method for comparing the strength of theories which has already been mentioned in Section 4.1.

Definition 124. Let Th, Th$'$ be any two theories. We say that Th$'$ has super-polynomial **speed-up** over Th if there exists a sequence of sentences $(\phi_i)_{i<\omega}$ such that for any i, the sentence ϕ_i is provable both in Th and Th$'$ and the function

$$f(i) = \frac{|\phi_i|_{\mathrm{Th}}}{|\phi_i|_{\mathrm{Th}'}}$$

cannot be dominated by any polynomial. We say that Th$'$ has super-polynomial speed-up over Th **relative to** a language $\mathscr{L}$ if the above holds for some sentences ϕ_i in the language $\mathscr{L}$.

[2] Actually, for a good notion of a size of proof we should either assume that some variables are written down with more than one symbol (as we have to write down their subscripts) or define the size of proofs simply as *Gödel codes of proofs* which are, after all, just some natural numbers. However, the latter definition would require us first to analyse in more detail how coding is carried out and, as a matter of fact, the slight inaccuracy in our definition will not really matter, since we are only concerned with *large* speed-ups (superpolynomial and above) and such speed-up relations are not affected by counting the size of variables which we may assume to change the size of proofs by at most polynomial factor.

Of course, there is nothing in definition of super-polynomial speed-up that prevents us from extending this definition to other classes of functions. One theory can have, e.g., polynomial, super-computable or super-exponential speed-up over another one with the obvious meaning which would be slightly awkward to define in full generality.

The study of speed-ups provides us with a very concrete way to compare theories. It measures how *efficient* a given theory is in proving theorems compared to another. If Th, Th$'$ are two theories of truth, by

$$\mathrm{Th} <_{\mathrm{sp}} \mathrm{Th}'$$

we mean that Th$'$ has super-polynomial speed-up over Th relatively to the language $\mathscr{L}_{\mathrm{PA}}$ of PA. Again, we will use symbols $=_{\mathrm{sp}}$, $\geq_{\mathrm{sp}}$ with their obvious meaning.

There is one easy way of ensuring that for two given theories of truth, Th_1, Th_2, the theory Th_2 has *at most* polynomial speed-up over Th_1. Namely, we can sometimes see how to directly rewrite proofs in Th_1 to proofs in Th_2. If Th_1 is finitely axiomatised, this may be enough to ensure that there is no super-polynomial speed-up between Th_1 and Th_2.

Proposition 125. *Suppose that* Th_1 *and* Th_2 *are two truth theories. Suppose that* Th_1 *is axiomatised by* PA *and finitely many truth-theoretic axioms and that it is relatively definable in* Th_2*. Then* Th_1 *does not have a super-polynomial speed-up over* Th_2*.*

Proof. We know that there exists a formula ϕ such that for any axiom α of Th_1,

$$\mathrm{Th}_2 \vdash \alpha[\phi/T].$$

Since there are only finitely many axioms α of Th_1 which contain the predicate T, there exists a constant C such that for any axiom α of Th_1,

$$|\alpha[\phi/T]|_{\mathrm{Th}_2} \leq C|\alpha|_{\mathrm{Th}_1}.$$

Without loss of generality, we can assume that the number of symbols in ϕ is not greater than C. Now, take any proof d in Th_1. Let d' be a sequence of formulae resulting from d by replacing every occurrence of T with ϕ (possibly renaming variables so as to avoid clashes) and appending to every axiom α of Th_1 a proof of the sentence $\alpha[\phi/T]$. Now, the number of symbols in d' is, at most, C times greater than the number of symbols in d which proves that $\mathrm{Th}_1 \leq_{\mathrm{sp}} \mathrm{Th}_2$. □

The simple argument applied above may sometimes be modified to handle the case when Th_1 is not finitely axiomatised. We can hope for an analogous proof whenever axioms of Th_1 may be interpreted in Th_2 in a *sufficiently uniform way*. This is a very vague description. An example of what this could mean is the following result, already mentioned in Section 2.1 in the remarks following Theorem 60. It has been originally proved in [Fischer, 2014], Theorem 2.

Proposition 126. UTB *does not have a super-polynomial speed-up over* PA*.*

It is not so easy to see how can we show that one theory *does* have a speed-up over another theory. This requires rather careful control on the size of *minimal* proofs in both theories. We should find some family of sentences which can be shown not to have relatively short proofs in a given theory. Luckily, there is a very useful and quite general result to this effect. We first need some notation:

Definition 127. Let Th be a theory whose language contains $\mathscr{L}_{\mathrm{PA}}$. We say that a unary formula $I(x)$ is a **cut** in Th if:

1. $\mathrm{Th} \vdash I(0)$.
2. $\forall x \Big(I(x) \to I(x+1) \Big)$.
3. $\forall x \forall y \Big(y < x \wedge I(x) \to I(y) \Big)$.

Now we may quote a result from [Fischer, 2014], Theorem 7, which is an application of techniques developed by Pudlák.

Theorem 128. *Let* Th *be a theory extending* PA*, such that for some unary formula* $I(x)$*,*

$$\forall x \Big(I(x) \to \neg \mathrm{Pr}_{\mathrm{PA}}(x, \ulcorner 0 = 1 \urcorner) \Big).$$

Then Th *has super-exponential speed-up over* PA*.*

Recall Definition 107: we say that Th_1 has super-exponential speed-up over Th_2 if there exists an infinite sequence (ϕ_i) of formulae provable both in Th_1 and Th_2 such that the function

$$f(i) = \frac{|\phi_i|_{\mathrm{Th}_1}}{|\phi_i|_{\mathrm{Th}_2}}$$

cannot be dominated by any function which can be obtained from exponential and constant functions by addition, multiplication and composition.

From this theorem, the following result can be deduced which has been first published in [Fischer, 2014], Theorem 9.

Theorem 129. $\mathrm{PT}^- +$ *INTtot has super-exponential (thus super-polynomial) speed-up over* PA*.*

We obtain another very natural division line between weak and strong theories of truth when we ask which of them enjoy significant speed-up over PA. As one can see, this division is somewhat orthogonal to the division obtained by considering semantic conservativity of truth theories, since a model-theoretically strong theory UTB can have no significant speed-up, as opposed to a fairly weak theory $\mathrm{PT}^- +$ INTtot. A natural question may be asked as to whether the classical compositional theory of truth CT^- which is syntactically conservative and model-theoretically strong allows to prove theorems more speedily than PA. However, according to an unpublished theorem of Enayat, Kaufmann, Visser, and Łełyk, this turns out not to be the case.

Theorem 130. CT^- *does not have a super-polynomial speed-up over* PA*.*

From the above result, Theorem 129, and Proposition 125, we may conclude that $\mathrm{PT}^- +$ INTtot is not relatively interpretable in CT^-, thus proving Proposition 121.

5.3. Interpretation of the results

In the introduction, we have outlined the philosophical debate which is the background for the study of strength of truth theories. Now let us summarise what conclusions may be drawn from our research which are relevant for the study of deflationism.

5.3.1. Proof-theoretic results

One group of the results presented in this thesis seems clearly relevant to the evaluation of the Shapiro-style conservativeness argument presented in our introduction, i.e. the argument assuming that the accurate truth theory should satisfy CT and concluding that it cannot be conservative. Let us recall that one possible way of strengthening this argument is to provide more examples of theories of truth which are not syntactically conservative over PA, but which are axiomatised with principles plausibly describing the natural truth predicate, especially when these principles are of purely truth-theoretic character and *prima facie* do not rely on our understanding of natural numbers. There was an objection raised by Field against Shapiro's original argument in [Field, 1999] that induction axioms for truth predicate do not necessarily fit the bill. In this context, it would be desirable to see some different examples of non-conservative truth theories whose axioms are motivated only by considerations about the notion of truth itself.

One such theory has been proposed by Cieśliński in [Cieśliński, 2010a]. The theory in question was $\mathrm{CT}^- + \mathrm{RP}$, whose axioms say that truth is compositional and closed under derivations in propositional logic. Arguably, its axioms are not dependent on any additional insights about natural numbers. However, the proof that this theory is not conservative was actually showing that it entails CT_0. Unfortunately, the result that the latter theory is not syntactically conservative was published in the literature with a gap in the proof. Therefore, to make Cieśliński's finding directly applicable to the conservativeness debate, one had to prove that CT_0 is a proof-theoretically strong theory. This is exactly what we have done in our thesis, in Theorem 63.

It is not clear whether all the characterisations of CT_0 provided by Cieśliński, Enayat, and Łełyk described in Section 2.2 provide additional support to Shapiro's conservativeness argument. Some of them—namely RFO, RP and SFO—seem to have purely truth-theoretic character because they describe interactions of truth with logic. The others—SPA and DC + INT—seem to mix truth-theoretic considerations with those concerning natural numbers because they basically express adequacy of PA. Therefore, they might be subject to similar objections as the original use of induction principles in Shapiro's argument.

The other relevant theories in this context are $\mathrm{CT}^- + \mathrm{DC}$ and $\mathrm{CT}^- + \mathrm{SP}$. They both seem to be very truth-theoretic and not really dependent on arithmetical considerations. Hence, if they turn out to be nonconservative over PA, then they could be used to produce other Shapiro-like arguments against deflationism.[3]

[3]Recently, Fedor Pakhomov has shown that $\mathrm{CT}^- + \mathrm{DC}$ is not conservative over PA and, in fact, shares arithmetical consequences with CT_0.

One can think of another closely related application of our results to support Shapiro's argument. Namely, Field raised an objection that non-conservativeness proof of CT employs induction axioms for the truth predicate, which he claims to depend more on our assumptions about natural numbers rather than on the notion of truth.

Note however that by Corollary 84, the theory $\mathrm{CT}^- + \mathrm{DC} + \mathrm{INT}$ is not syntactically conservative over PA, contrary to $\mathrm{CT}^- + \mathrm{INT}$. One can argue that this blocks Field's objection that truth-theoretic principles themselves are not responsible for the strength of the truth theory, but it is rather some new, essentially arithmetical axioms which we add alongside the truth theoretic ones. This objection seems to fail in the context of $\mathrm{CT}^- + \mathrm{INT}$, since CT^- with the modicum of induction provided by the principle INT is by itself no stronger than PA. It is only upon adding disjunctive correctness DC that we gain additional proof-theoretic strength.

It is not entirely clear whether this is really a good response to Field's argument because to use the strength of DC, we still used some form of induction. Even if this amount of induction is by itself weak, our result does not satisfy Field's requirement that the Shapiro-style conservativeness argument against deflationism be carried out using purely truth-theoretic principles.

In fact, if this response to Field's criticism is valid, then one can formulate a much simpler one. Namely, we should note that the scheme of induction for the sentences containing the truth predicate is also conservative over PA, unless induction is coupled with some axioms to the effect that the truth predicate is, in fact, compositional. Only then one can prove all the arithmetical consequences of CT extending PA. In a way, already in this very simple scenario, it is only upon adding truth-theoretic principles that we are able to *use* induction for the truth predicate in a nontrivial way.[4]

5.3.2. Other notions of strength

As we have already seen, proof-theoretic results on truth theories may be rather directly applied to the debate on deflationism. In this thesis, we have considered some other notions of the strength, such as semantic strength, strength with respect to relative definability and strength with respect to the efficiency (i.e. length) of proofs. Let us briefly discuss what they can possibly bring to the discussion of deflationism.

Let us first discuss the notions of strength given by relative definability and the comparison of lengths of proofs. The first of them has been introduced to capture the notion of the *conceptual strength* of a theory: if Th_1 can relatively define Th_2, then it can capture what it means to be true in the sense of Th_2. We can think of no obvious applications of relative definability results to the debate on deflationary theory of truth.

If relative definability were our measure of strength, then TB^- would already be stronger than PA, simply by Tarski's theorem. It seems that there is really no room for the claim that truth is a weak notion in the sense captured by relative definability. On the other hand, this relation seems very natural and well-motivated and it is very likely that it will find its applications in future debates.

[4]Note that due to Pakhomov's result that $\mathrm{CT}^- + \mathrm{DC}$ is as strong as CT_0, we know that the strength of $\mathrm{CT}^- + \mathrm{DC} + \mathrm{INT}$ does not depend at all on the principle of internal induction.

The notion of speed-up is meant to capture the *usefulness* of a given theory of truth. A theory of truth Th_1 has speed-up over (the arithmetical part of) a theory of truth Th_2 if Th_1 can prove some arithmetical sentences *faster* than Th_2. Again, this notion is undoubtedly very natural. Unfortunately, it is not entirely clear whether it is really applicable to the debate on the deflationary theory of truth. Superficially, this notion seems very strongly related to typical claims of deflationists because they often stress precisely the usefulness of truth predicate.

However, the truth predicate is supposed to be useful in that it brings additional expressive power, most notably in that it allows us to handle infinite conjunctions and disjunctions. This notion of usefulness is unfortunately not really linked to the one produced by comparison of lengths of proofs. That being said, let us note that speed-up is a very natural phenomenon to be investigated and possibly the research on how adding a truth predicate affects the lengths of proofs when compared to the base theory will find its applications in the debate on deflationism.

Let us finally discuss the results on semantic strength. First, note that model-theoretic considerations are a very handy tool for obtaining negative results about relative definability (as in Corollary 119). Therefore, even if one finds applications of semantic conservativity to the debate on deflationism rather suspicious, then the investigation of models of truth theories may still be motivated by being applied to the study of relative definability, which is a purely syntactic relation.

Now, the most obvious way to apply the results on semantic strength to the debate on deflationism would probably be to modify the conservativeness argument of Shapiro so that it employs the notion of model-theoretic conservativeness instead of the notion of proof-theoretic conservativeness.

In the original Shapiro-style conservativeness argument, one argues that if truth is an insubstantial notion, then employing this notion should not permit us to draw any new consequences concerning the facts which are expressible in the base theory. In the modified version, one could insist that if truth is an insubstantial notion, then this does not merely mean that we cannot *prove* new things using this notion. The fact that we can attach to our theory a truth predicate satisfying certain axioms should not make any difference at all to how the world looks.

Basing the conservativeness argument on the notion of semantic strength should make this argument substantially more immune to criticism as to whether the choice of axioms for the truth predicate is really innocent and justified. Namely, as we have already seen in Chapter 3, already such theories as $\mathrm{TB}, \mathrm{UTB}$ or CT^- turn out to be semantically non-conservative over PA (see Propositions 90, 92, and Theorem 98, respectively). As we have already mentioned, Field criticised the use of CT in the conservativeness argument by Shapiro, since he deemed its axioms involving the truth predicate to be a mix of truth-theoretic and arithmetic principles.

The fact that CT^- is not semantically conservative over PA may be possibly a good response to this kind of criticism. More generally, one natural way of criticising conservativeness arguments is to claim that the truth-theoretic axioms postulated in these arguments are not the ones the deflationist would accept. Therefore, the exploration of semantic strength of the theories of truth is possibly of some use because it is much

easier to obtain examples of theories which are semantically non-conservative than examples of proof-theoretic non-conservativity.

Nonetheless, there remain two major problems with this strategy. The first one is rather obvious: Why should semantic conservativeness of truth theories be the correct explication of the deflationary theory of truth? We have to admit honestly that we do not have a good answer to this question. On the other hand, it is also not clear to us why *syntactic* conservativeness should be the right explication of the claim that truth is a "metaphysically thin" notion. Obviously, there is some intuition that notions which are merely logical devices should not enable us to draw any new conclusions about the concrete facts, but the claim that they should not impose any additional structural conditions as to how the world looks like *also* seems intuitive.

The other major problem is the one discussed already in the introduction. Strictly speaking, even if one moves to semantic conservativeness, then the two main claims of deflationism are still mutually consistent. Recall that the claims in question are as follows:

1. The content of the truth predicate is encapsulated in Tarski's biconditionals.

2. Truth is an insubstantial notion.

Even if we are ready to explicate the negative claim so that it entails that truth theory is semantically conservative over its base theory, then the deflationist can still maintain both claims. As we have seen in Proposition 89, UTB^-, the theory of uniform Tarski's biconditionals without induction is semantically conservative over PA. Hence, if the deflationist does not have an independent reason to hold that truth predicate must satisfy full induction scheme, then we cannot achieve any contradiction between these claims.

We cannot even require the deflationist to *deny* that the truth predicate is compositional or fully inductive. Even if we have no guarantee that the truth predicate is compositional or that it satisfies induction, because that would impose some nontrivial conditions on the structure of the world, then it might still turn out that the world *does* satisfy these conditions anyway. Hence, the deflationist is at the very best forced to leave open the possibility that the truth predicate need not satisfy anything besides the basic disquotational scheme. However, the claim that we have no *guarantee* that we can attach a compositional or an inductive truth predicate to our theory is not very strong.

The semantic nonconservativity of TB and UTB—theories axiomatised with Tarski's biconditionals and full induction for the truth predicate—is probably slightly more problematic because the deflationist is now forced to claim that *induction* is not guaranteed to be satisfied by the truth predicate. On the other hand, as we have already mentioned in this section, there does exist a deflationist who has precisely claimed that the induction scheme for the truth predicate does not follow from the concept of truth itself. Hence, nonconservativity of TB and UTB is by no means an ultimate argument against the deflationary theory of truth.

Bibliography

Cantini, A. (1990). A theory of formal truth as strong as ID_1. *The Journal of Symbolic Logic*, 55(1):244–259.

Cieśliński, C. (2010a). Deflationary truth and pathologies. *The Journal of Philosophical Logic*, 39(3):325–337.

Cieśliński, C. (2010b). Truth, conservativeness and provability. *Mind*, 119(474):409–422.

Cieśliński, C. (2011). T-equivalences for positive sentences. *The Review of Symbolic Logic*, 4(119):319–325.

Cieśliński, C. (2015a). The innocence of truth. *Dialectica*, 69(1):61–85.

Cieśliński, C. (2015b). Typed and untyped disquotational truth. In Achourioti, T., Galinon, H., Martínez Fernández, J., and Fujimoto, K., editors, *Unifying the Philosophy of Truth*, pages 307–320. Springer.

Cieśliński, C. (2017). *The Epistemic Lightness of Truth. Deflationism and its Logic*. Cambridge University Press.

Cieśliński, C., Łełyk, M., and Wcisło, B. (2017). Models of PT^- with internal induction for total formulae. *The Review of Symbolic Logic*, 10(1):187–202.

Enayat, A. and Visser, A. (2015). New constructions of satisfaction classes. In Achourioti, T., Galinon, H., Martínez Fernández, J., and Fujimoto, K., editors, *Unifying the Philosophy of Truth*, pages 321–325. Springer.

Feferman, S. (1991). Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56(1):1–49.

Field, H. (1999). Deflating the conservativeness argument. *The Journal of Philosophy*, 96(10):533–540.

Fischer, M. (2009). Minimal truth and interpretability. *The Review of Symbolic Logic*, 2(4):799–815.

Fischer, M. (2014). Truth and speed-up. *The Review of Symbolic Logic*, 7(2):319–340.

Fischer, M. (2015). Deflationism and instrumentalism. In Fujimoto, K., Fernández, J. M., Galinon, H., and Achourioti, T., editors, *Unifying the Philosophy of Truth*, pages 293–306. Springer Netherlands.

Fischer, M. and Horsten, L. (2015). The expressive power of truth. *The Review of Symbolic Logic*, 8(2):345–369.

Franzen, T. (2004). *Inexhaustibility: an Inexhaustive Treatment*. A K Peters/CRC Press.

Fujimoto, K. (2010). Relative truth definability in axiomatic truth theories. *Bulletin of Symbolic Logic*, 16(3):305–344.

Hájek, P. and Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Springer.

Halbach, V. (2009). Reducing compositional to disquotational truth. *The Review of Symbolic Logic*, 2(4):786–798.

Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge University Press.

Horsten, L. (1995). The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In Cartois, P., editor, *The Many Problems of Realism (Studies in the General Philosophy of Science: Volume 3)*, pages 173–187. Tilberg University Press.

Horsten, L. (2009). Levity. *Mind*, 118(471):555–899.

Horsten, L. and Leigh, G. (2017). Truth is simple. *Mind*, 126(501):195–232.

Horwich, P. (2001). A defense of minimalism. *Synthese*, 126(1/2):149–165.

Kaufmann, M. (1977). A rather classless model. *Proceedings of American Mathematical Society*, 62(2):330–333.

Kaufmann, M. and Schmerl, J. (1987). Remarks on weak notions of saturation in models of Peano Arithmetic. *Journal of Symbolic Logic*, 52(1):129–148.

Kaye, R. (1991). *Models of Peano Arithmetic*. Oxford: Clarendon Press.

Ketland, J. (1999). Deflationism and Tarski's paradise. *Mind*, 108(429):69–94.

Kossak, R. and Schmerl, J. (2006). *The Structure of Models of Peano Arithmetic*. Oxford Science Publications.

Kotlarski, H. (1986). Bounded induction and satisfaction classes. *Zeitschrift für matematische Logik und Grundlagen der Mathematik*, 32(31–34):531–544.

Kotlarski, H. (1991). Full satisfaction classes: A survey. *Notre Dame Journal of Formal Logic*, 32(4):573–579.

Kotlarski, H., Krajewski, S., and Lachlan, A. (1981). Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin*, 24:283–93.

Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy*, 72(19):690–716.

Lachlan, A. H. (1981). Full satisfaction classes and recursive saturation. *Canadian Mathmematical Bulletin*, 24:295–297.

Leigh, G. (2015). Conservativity for theories of compositional truth via cut elimination. *The Journal of Symbolic Logic*, 80(3):845–865.

Łełyk, M. and Wcisło, B. (2017a). Models of weak theories of truth. *Archive for Mathematical Logic*, 56(5):453–474.

Łełyk, M. and Wcisło, B. (2017b). Notes on bounded induction for the compositional truth predicate. *The Review of Symbolic Logic*, 10(3):455–480.

Pudlák, P. (1986). On the length of proofs of finitistic consistency statements in first order theories. In Paris, J. B., Wilkie, A., and Wilmers, G., editors, *Logic Colloquium 84*.

Pudlák, P. (1998). The lengths of proofs. In Buss, S. R., editor, *Handbook of Proof Theory*, pages 547–642. Elsevier.

Quine, W. V. O. (1946). Concatenation as a basis for arithmetic. *The Journal of Symbolic Logic*, 11(4):105–114.

Schmerl, J. (1981). Recursively saturated, rather classless models of Peano arithmetic. In Lermal, M., Schmerl, J., and Soare, R., editors, *Logic Year 1979–1980: The University of Connecticut*, pages 268–282. Springer.

Shapiro, S. (1998). Proof and truth: Through thick and thin. *The Journal of Philosophy*, 95(10):493–521.

Shelah, S. (1978). Models with second order properties II. Trees with no undefined branches. *Annals of Mathematical Logic*, 14(1):73–87.

Smith, S. T. (1989). Nonstandard definability. *Annals of Pure and Applied Logic*, 42(1):21–43.

Tarski, A. (1995). Pojęcie prawdy w językach nauk dedukcyjnych. In Zygmunt, J., editor, *Pisma Logiczno-Filozoficzne*, volume 1. Państwowe Wydawnictwo Naukowe.