Abstract

Title: Axiomatic Studies of Truth

Name: Kentaro Fujimoto

College: Merton College

Degree: DPhil in Philosophy

Date of Submission: December 22, 2010.

Approximate Number of Words: 72000 words (both including and excluding citations)

In contemporary formal theory of truth, model-theoretic and non-classical approaches have been dominant. I rather pursue the so-called classical axiomatic approaches toward truth and my dissertation begins by arguing for the classical axiomatic approach and against the others. The classical axiomatic approach inevitably leads to abandonment of the naïve conception of truth and revision of the basic principles of truth derived from that naïve conception such as the full T-schema. In the absence of the general guiding principles based on that naïve conception, we need to conduct tedious but down-to-earth 'field works' of various theories of truth by examining and comparing them from various points of view in searching for satisfactory theories of truth. As such attempt, I raise two new criteria for comparison of truth theories, make a proof-theoretic study of them in connection to the foundation of mathematics.

Contents

1 The Classical Axiomatic Approach **5**

1.1 Definitional Approach vs Non-definitional Approach 5

1.2 Semantic Approach vs Axiomatic Approach 16

1.3 Revision of Logic 22

1.3.1 Moh Shaw-Kwei's Paradox 26

1.3.2 Hajek-Paris-Shepherdson Paradox 35

1.4 The Classical Axiomatic Approach 40

2 List of Systems of Truth **43**

2.1 Notational Preliminaries 43

2.2 Disquotational systems of truth I 47

2.3 Systems of Typed Truth 50

2.4 Systems of iterative compositional truth I 54

2.5 Disquotational systems of Truth II 61

2.6 Systems of iterative compositional truth 2 62

2.7 Systems of Determinate Truth 65

2.8 Systems of Symmetric Truth 68

2.9 Systems of iterative non-compositional truth 75

2.10 Schematic Reflective Closures 76

3 Relative Truth Definability **78**

3.1 Comparing Conceptual Aspects of Axiomatic Truth 78

3.2 Relative Truth Definability 83

3.2.1 Mathematical Properties of Truth-Definability 88

3.3 Digression – Proof-Theoretic Ordinal – 90

3.4 Several Results on Truth Definability 92

3.5 Some Applications 102

3.6 Concluding Remarks of Chapter 3 106

4 The Inner Theory of Truth Systems **109**

4.1 The Notion of Inner Theory 109

4.2 Systems of Finitely Iterated Self-Applicable Truth 114

4.3 Semantics of iterative compositional theories 118

4.4 Inner Theory Comparison via Semantics 125

4.5 Proof theory for finitely iterated self-applicable truths 129

4.5.1 Iterated fixed-points and self-applicable truths 130

4.5.2 Semi-Formal Systems for $\widehat{\mathsf{ID}}_n$ 132

4.5.3 Asymmetric Interpretation 136

4.6 Inner theory comparison via semi-formal systems 143

4.6.1 Systems with and without Cons 143

4.6.2 Comparison of different self-applicable truths 150

4.7 Summary of Chapter 4 151

5 Truth Progressions **153**

5.1 Background 153

5.2 Notation for Transfinitely Iterated Truths 156
5.3 Kripke-Feferman Truth 159
5.4 Consistency and Completeness Axioms 162
5.5 Positive Uniform T-biconditionals 175
5.6 Kripke-Feferman Truth with Weak Kleene Logic 177
5.7 Determinate Truth and Feferman Logic 184
5.8 Cantini's VF-Truth 187
5.8.1 The Systems $\mathsf{VF}_{\prec}$ and $\mathsf{Aut}(\mathsf{VF})$ 188
5.8.2 Lower Bound 190
5.8.3 Upper Bound 195
5.9 Summary and Concluding Remarks of Chapter 5 202

6 Further Prospects **204**

Chapter 1

The Classical Axiomatic Approach

In my thesis, I concentrate on so-called formal theories of truth. Among various types of formal theories of truth, I choose one type of theories that I call 'classical axiomatic' theories of truth. In this opening chapter, I will classify formal theories of truth into some categories, compare the advantages and disadvantages of them, and then explain why I prefer the classical axiomatic approach to the others.

1.1 Definitional Approach vs Non-definitional Approach

Over a long period of time, a huge variety of theoretical analyses of the notion of truth have been attempted. In the contemporary debates on truth, after the development of modern symbolic logic, a distinction is particularly made between philosophical and formal theories of truth. If this distinction is granted, my principal interest of the present thesis lies in the latter, although they bear a quite close and reciprocal connection to each other.

Notoriously, the notion of truth bears a paradoxical aspect. Due to this aspect, we are always faced with the danger of inconsistency in any kind of analyses of truth. The most famous and basic

example is the Liar paradox. One formulation of this paradox is:

The statement (1.1) is false. (1.1)

This example might look unnatural. Alternatively, we may consider the following situation: a stranger comes to you on the street, he mistakes you for somebody else, and then suddenly says to you 'I never tell you the truth'; but he is a stranger to you and you have never met him before; namely, the above is the first utterance made by him toward you. Can this utterance either true or false? This 'Liar' problem is not easily solved at all. Admitting statements neither true nor false cannot be an immediate solution because of the so-called strengthened Liar paradox. Restricting self-reference doesn't straightforwardly lead us to a solution either because of other paradoxes with no appeal to self-reference such as Yablo's paradox [86].

This paradoxical phenomenon related to the notion of truth is indeed a serious obstacle for theoretical analyses of this notion. Paul Horwich proposed a distinction between 'the theory of the concept of truth' and 'the theory of truth itself'. According to him, 'the former purports to specify the conditions in which someone uses the word 'true' with a certain meaning ([39, p.252])' and the latter purports to specify 'principles about the *property* of truth on the basis of which all the facts about truth are to be explained ([39, p.253]).' For the former, Horwich insists, the inconsistency arising from the Liar paradox may not be an overwhelming difficulty. McGee also pointed out 'it is possible to use the word 'true' coherently without possessing a coherent theory of truth ([58, p.3])'. It might well be supposed that a certain condition constitutes our practical mastery of the word 'true' but leads to inconsistency at the same time. In the present thesis, however, I am principally interested in 'the theory of truth itself' in the terminology of Horwich, and thus inconsistency is the worst consequence to be avoided in the study of this type of theories of truth.

In the formal approach to the notion of truth, we facilitate the tools of contemporary formal

logic. One of the principal benefits of the use of formal logic concerns the aforementioned problem of inconsistency. By means of formal logic, we can investigate whether a given system of truth is consistent or not. Formal tools provides us with a framework in which we can measure how safe a given theory is and make it more explicit what may cause a contradiction and how we can avoid it.

Then, what are formal theories of truth like? There are various kinds of such theories. Let us try to classify them into some categories. They first branch out into two categories: definitional or reductionist theories, and non-definitional or non-reductionist theories. The former group aims at giving a definition of truth and thereby explaining away it in terms of other more basic and non-semantical notions. There can be considered two types of definitional approaches. The one tries to define the truth of a given language $\mathcal{L}$ in the terminology of $\mathcal{L}$ itself; the other aims at reducing it into some terminology possibly not possessed by $\mathcal{L}$. However, both types of definitional approaches are no longer a principal concern of truth theorists.

Only a few contemporary philosophers and logicians would choose the former stronger form of definitional approaches of defining the truth of $\mathcal{L}$ *within* $\mathcal{L}$. This common renunciation comes from the well-known *undefinability theorem* of Tarski. Tarski [82] proved that a definition of truth is impossible under fairly general assumptions. Let $\mathcal{L}$ be the language whose truth we want to consider. If the truth of $\mathcal{L}$ be definable within $\mathcal{L}$, we have an $\mathcal{L}$-predicate Φ of $\mathcal{L}$-sentences which defines the truth of $\mathcal{L}$. Since the truth predicate is a predicate of sentences, we assume $\mathcal{L}$ contains names of $\mathcal{L}$-sentences and let $\ulcorner\sigma\urcorner$ be the name of a sentence σ. Tarski showed that, under fairly moderate conditions on $\mathcal{L}$ and the background theory over $\mathcal{L}$, there is no predicate Φ such that, for all $\mathcal{L}$-sentence σ,

$$\Phi(\ulcorner\sigma\urcorner) \text{ if and only if } \sigma;$$

this schema is called Tarski's *T-schema* and each instance of the T-schema is called *T-biconditional*.

The famous Tarski's convention T claims that

> A predicate Φ is materially adequate definition of the truth of $\mathcal{L}$ iff (α) the T-schema holds and (β) if $\Phi(x)$ then x is equal to $\ulcorner\sigma\urcorner$ for some $\mathcal{L}$-sentence σ.

Tarski seems to be claiming that these (α) and (β) are *necessary* and *sufficient* conditions for a predicate Φ to be a definition of truth. It is still arguable whether they are sufficient on the one hand; but it appears to be a common view among philosophers and logicians that they are necessary. Hence, most of philosophers and logicians would renounce any definitional approach of this stronger form; at least, they do not start their study of truth by assuming that such a definition of truth is possible.

In contrast, the weaker form of the definitional approach will raise little concern either, simply because the standard and commonly accepted definition of truth is already in our hands: i.e., most philosophers and logicians would agree that Tarski's original definition of the truth of a language $\mathcal{L}$ by means of the mathematical resource outside of $\mathcal{L}$ is successful and acceptable to a great degree.[1] Though his original setting was based on a slightly old-fashioned theory of types (or calculus of classes), his methodology can be easily carried over to contemporary settings. For example, we can carry out Tarski's construction of the truth of the language $\mathcal{L}_{\mathsf{PA}}$ of first-order arithmetic within the second-order system ATR_0 of arithmetical transfinite recursion; we simulate his construction for the language $\mathcal{L}_{\in}$ of first-order set theory within the second-order system NBG_ω of ω elementary recursion; under the standard translation of the language of arithmetic in that of set theory, the Tarskian definition of the truth of $\mathcal{L}_{\in}$ can be given in a fairly weak set theory, say, $\mathsf{KP}\omega$; in these three examples of the Tarskian definition of truth, ATR_0, NBG_ω and $\mathsf{KP}\omega$ plays the role of the *metatheory* or *metasystem* (in what follows, though it is not standard, I will use the word 'metasystem' for technical reasons), their languages are called *metalanguage* and the languages

[1] By 'Tarski's theory', I mean his theory illustrated in the first half of §3 and in §4 of [82], and I exclude another theory of his developed in [82, pp.199–209]. It is the former one that we call 'Tarski's theory of truth' today.

$\mathcal{L}_{\mathsf{PA}}$, $\mathcal{L}_{\in}$ and $\mathcal{L}_{\mathsf{PA}}$ respectively correspond to the object-language. Hence, as for this weaker form of the definitional approach, the problem is not on whether or not definition is possible. Rather the problem is that such a definition bears little significance for our purposes.

Remark 1. It is sometimes wrongly thought that Tarski's definition of truth in his [82] just amounts to giving the ordinary classical model-theoretic structure $\mathfrak{M}$, which is also introduced and defined by Tarski himself and his student Vaught [83], of the language in question. Though their definitions are technically similar, they should be clearly distinguished.

For the sake of the subsequent arguments, we slightly digress for a while and look more closely at the Tarskian definition. We must start by fixing a metasystem Q over a certain language $\mathcal{L}_{\mathsf{Q}}$. Let $\mathcal{L}$ be the target language the truth of which is to be defined. In general, there need not be any particular connexion between $\mathcal{L}_{\mathsf{Q}}$ and $\mathcal{L}$; one may be a proper sublanguage of the other; they may be identical; they may be entirely disjoint; or they may partially overlap. However, it is necessary for the Tarskian definition of the truth of $\mathcal{L}$ that $\mathcal{L}$ can be properly translated into $\mathcal{L}_{\mathsf{Q}}$. Thereby, the definition of the truth of $\mathcal{L}$ is carried out within Q as the definition of the truth of the translation of $\mathcal{L}$ in $\mathcal{L}_{\mathsf{Q}}$. Then, according to Tarski's theorem, only when $\mathcal{L}$ (or, equivalently, its translation in $\mathcal{L}_{\mathsf{Q}}$) is essentially poorer than $\mathcal{L}_{\mathsf{Q}}$, the truth of $\mathcal{L}$ is definable in Q. Now, let us recall that Tarski began his arguments by explicitly stating that

> The question how a certain concept is to be definable is correctly formulated only if a list is given of the terms by means of which the required definition is to be constructed. If the definition is to fulfil its proper task, the sense of the terms in this list must admit of no doubt. [82, p.152–153]

That is, the sense of $\mathcal{L}_{\mathsf{Q}}$-terms must be given and fixed in advance before the definition of truth for $\mathcal{L}$.[2] Consequently, each $\mathcal{L}$-term has the same meaning as its translation in $\mathcal{L}_{\mathsf{Q}}$ which is already given by virtue of Q. Hence, in order to talk about the translatability of $\mathcal{L}$ in $\mathcal{L}_{\mathsf{Q}}$ (with the background

[2]It is highly debatable how we can provide a term with its meaning. Some might only require some axioms and inference rules; other might require more. But I don't get into this issue and let us assume that the sense of $\mathcal{L}_{\mathsf{Q}}$-terms is given by a formal axiomatic system Q.

system Q), the sense of $\mathcal{L}$-terms is also to be given. Now consider the case where we define the truth of the language $\mathcal{L}_{\in}$ within a second-order set theory such as Morse-Kelly impredicative class theory MK ($\supset \mathsf{NBG}_{\omega}$) over the language $\mathcal{L}_2$ ($\supset \mathcal{L}_{\in}$) of second-order set theory. Then, the sense of each term of the target language $\mathcal{L}_{\in}$ is given within MK. Let ZFA be ZF plus Aczel's anti-foundation axiom. ZFA is a system over the same language $\mathcal{L}_{\in}$ but not a subsystem of MK. It is sometimes argued that the conception of sets is different between ZF and ZFA; the former is based on, say, the iterative conception of sets, and the latter is based on the so-called graph conception of sets; therefore, we might well say that the sense of '$\in$' is different between ZF and ZFA. That is to say, $\mathcal{L}_{\in}$ with the background system ZF and $\mathcal{L}_{\in}$ with ZFA are different languages with different senses; the mere syntactical coincidence of their vocabulary does not necessarily entail that each term has the same meaning as each other. In the above example, this point is put in the following way: MK has defined the truth of the language $\mathcal{L}_{\in}$ with some meaning but has not defined the truth of the syntactically identical language $\mathcal{L}_{\in}$ with another meaning. In summary, we cannot treat the object-language $\mathcal{L}$ simply as a syntactical collection of certain vocabulary; rather we need to treat it together with its background system by virtue of which the sense of each $\mathcal{L}$-term is given.

Keeping in mind the points developed in the last paragraph, we turn back to the methodological problems concerning the definitional approach. As is well-known, Tarski deduced from his aforementioned results the famous 'impossibility thesis' of truth one formulation of which says:

(Tars) The definition of truth of a language $\mathcal{L}$ is only possible in an essentially richer language $\mathcal{L}'$ (together with a sufficiently rich background system over $\mathcal{L}'$) than $\mathcal{L}$, but impossible in $\mathcal{L}$ itself.

One consequence of this thesis is crucial: if we stick to the definitional approach and require the truth predicate to be defined away, then no language $\mathcal{L}$ possesses its own truth predicate and only an 'essentially richer' language $\mathcal{L}'$ than $\mathcal{L}$ may possibly contain the truth predicate of $\mathcal{L}$. Hence, in particular, if our natural language is counted within the scope of the formal study of truth,

definitional theorists of truth must conclude that our natural language does not possess its own truth predicate, because vocabulary outside or beyond our natural language is no longer a part of our natural language of course and we cannot express them within our natural language. Let us recall Tarski's notion of 'universal language'. According to Tarski, a universal language is a language with 'adequate facilities for expressing everything that can be expressed at all, in any language whatsoever ([81, p.89]).' One of the conclusions Tarski deduced from the above (Tars) and his undefinability theorem is:

(Univ) We cannot define the truth of universal languages.

Since a natural language must be universal in the above sense (at least from the viewpoint of the speakers of that natural language), the impossibility of the truth definition for a natural language can be regarded as a corollary of Tarski's thesis (Univ).[3] Hence, a universal language cannot possess its own truth predicate in the definitional approach.

Not only are natural or universal languages to be excluded from the definitional approach. Tarski's construction is carried out within mathematics, and the theory he uses in the construction belongs to mathematics. The language of mathematics (or, mathematics simpliciter) is not universal in Tarski's original sense. However, trivially speaking, mathematics is closed under mathematical objects and reasonings: anything mathematical is subsumed in mathematics. Hence, mathematics constitutes, say, a 'closed totality' and thus we must conclude that we cannot define the truth of any language which includes that of the entire mathematics, unless we compromise to define the mathematical truth by means of non-mathematical tools which (probably) lack sufficient rigour. To

[3]The thesis (Univ) is made in [82]. Tarski later gave another similar thesis in [80]. He introduced the notion of a 'semantically closed' language; a language is called semantically closed if it contains 'in addition to its expressions, also the names of these expressions, as well as semantic terms such as the term 'true' referring to sentences of this language [80, p.123]'. Then, he claimed that

(Sem) We cannot define truth for semantically closed languages.

Then, it is natural to think that semantical closedness follows from universality. At any rate, the most basic thesis is that if a theory includes a reasonably rich theory of its own syntax, then it cannot define its own truth; then both (Univ) and (Sem) are regarded as a corollary of this thesis.

overcome this problem, one might adopt the so-called 'orthodox approach', in which the language of mathematics is conceived as layered into certain hierarchy of sublanguages, and the truth of one sublanguage is possessed by other languages higher than it in that hierarchy. However, it wouldn't help, because we couldn't express the statement 'there is at least one untrue mathematical theorem'; whether or not this statement is indeed true doesn't matter, but the problem is that we can not even express this statement. If one conceives mathematics as a static closed totality, this consequence should be rejected, unless she decides not to talk about the truth of the entire mathematics at all. It is also to be noted that the same type of 'orthodox approach' could be adopted for natural or universal languages but it would be refuted in a similar manner; consider the statements 'there is no sentence which is both true and false at the same time'; according to the current standpoint in question, this statement cannot bear a 'universal' or 'logical' character and the term 'no sentence' must be relativized to a certain initial fixed proper fragment of our language. In addition, Kripke's ingenious example on self-applicability [50] illustrates another defect of the 'orthodox approach' with respect to natural languages and it seems to me decisive and undefeatable.[4]

One might counterargue that mathematics is open-ended and does not constitute a static 'closed totality'; then, she might continue, whenever we try to define the truth of mathematics, the notion or realm of mathematics simply gets expanded. Although I myself have sympathy with the idea that mathematics is open-ended, this option shouldn't be taken for the very problem we are concerned with at any rate, simply because she has already expanded the notion or realm of mathematics in her attempt to define the truth of 'the former' mathematics and thus the resulting definition is no longer the definition of the truth of 'the new' mathematics. Nonetheless, this view gives us another perspective. If mathematics is open-ended and it never forms a closed totality, then it would be of little importance and significance to try to define the truth of the entire mathematics any longer;

[4]To my knowledge, the term 'orthodox approach' was first used by Kripke in [50] to denote this type of standpoint particularly for natural languages. It is not so clear if his criticism could have the same effect on the case of the language of mathematics, since self-applicability of the truth predicate might be dispensable for the case of mathematics.

consequently, the impossibility of the definition of the truth of mathematics would be no problem at all for definitional theorists of truth. A parallel arguments for natural language can be made: i.e., there is no totality of a natural language and thus no need to define the truth of the entire natural language.

Indeed, I myself am opt for this view of 'open-ended' mathematics. At least for our current knowledge, it seems too much to assume that the closed totality of mathematics is already given in front of us and we have already known exactly what constitutes mathematics. If such a fixed closed totality were given to us, we could concentrate on the fragments of mathematics whose truth is definable within that given totality of mathematics but would have to abandon any discourse concerning the truth of the fragments whose truth is not definable therein; consequently, we would have to anyhow abandon the definitional approach toward some parts of mathematics. Motivated by the 'open-ended' view, one may argue, for the definitional approach, that, whether or not Q is a system provisionally adopted as the representation of the entire mathematics, we could simply expand Q to a certain sufficiently stronger system Q' and then give the Tarskian definition of the truth of the language $\mathcal{L}_{\mathsf{Q}}$ of Q within Q'.[5] As a matter of fact, the motivation and background of this approach is resonant with those of the approach I myself will take. However, even if one adopts the above suggested approach, the definitional approach still bears two serious problems and it looks less versatile than my approach.

The first problem of the definitional approach in this approach concerns the self-applicability of truth: that is, we cannot therein deal with the self-applicable character of the notion of truth. This is simply because the truth predicate of a language $\mathcal{L}$ always belongs to the outside of $\mathcal{L}$ and thus

[5] I would like to emphasize again that Tarski's definition of truth is fairly satisfactory particularly for mathematical languages or any other languages, such as those of physics or chemistry, which do not intrinsically contain the truth predicate or other semantical predicates and that if the Tarskian definition of truth of a language $\mathcal{L}$ of our interest can be successfully given within an initially fixed metasystem Q, this is an almost ideal situation and we need not demand more, with the proviso that Q is acceptable system. However, we probably want to ask and talk about the truth of any given mathematical language and my interest of the present thesis is principally in the truth *simpliciter* of any language given to us for whatever reason, whether Q is given as the foundation of mathematics or a subject of a particular mathematics research.

it is meaningless and impossible to apply the truth predicate to a sentence with the truth predicate itself. Notice that this problem applies to any type definitional approach in general. This fact is a crucial defect particularly when we consider the truth of our natural language or any language with reasonably rich semantical terms. Formal theories of truth usually treat only mathematical languages, but if the ultimate goal of such formal study is to analyze our ordinary notion of truth in our natural language and to give an appropriate theory of truth for it, this defect is fatal and should not be ignored by any means.

The second problem is methodological one and comes from the fact that there is no unique procedure of expanding the initial (meta) system Q. It is to be observed that the Tarskian definition of truth of $\mathcal{L}_{\mathsf{Q}}$ within a stronger system S requires some redundant conditions on S which are not pertinent to the truth definition in their own. For example, as I have mentioned, we can carry out essentially the same construction of the truth predicate of $\mathcal{L}_{\in}$ in NBG_{ω} (and thus in MK *a fortiori*). However, as a matter of fact, we only need NBG to define a predicate which merely satisfies the T-schema for $\mathcal{L}_{\in}$ (cf. [62]); furthermore, within the system NBG plus ω-induction for Σ^1_1-formulae, we can define a predicate which satisfies all the semantical (and syntactical) properties concerning $\mathcal{L}_{\in}$ that Tarski used and mentioned as the consequence of his definition in [82]. Then we naturally ask which system we should choose to define the truth of $\mathcal{L}_{\in}$. There are at least three options to take:

(i°) Choose the weakest system among the systems in which we define a predicate satisfying the T-schema or a certain collections of the principles mentioned by Tarski.

(ii°) Choose the weakest system among the systems in which we can carry out the Tarskian definition.

(iii°) Choose whatever system as long as it achieves our purpose of carrying out the Tarskian definition of truth or satisfying a certain truth-theoretic principles.

First, the former two are faced with the problem of measuring the 'strength' of systems. To make the matter worse, even if we are given some appropriate measure of 'strength', it is not necessarily the case that the given measure well-orders all the relevant systems and thereby uniquely determines the 'weakest' one; in fact, most of the standard intertheoretic measures do not determine a unique such system. This problem might well be overcome simply by regarding all minimal systems according to a fixed measure as 'equivalent' and identifying them.

Second, as for the second option (ii°), the phrase 'can carry out the Tarskian definition' is fairly vague and it is difficult to give a general criterion to judge whether a given construction of a predicate is 'Tarskian' or not.

Third, as I have explained, in the Tarskian definition of truth, we apparently have to add redundantly strong assumption to the original system Q which has little direct connection to the notion of truth itself. If a certain supersystem S of Q is given for some reason independent of the truth definition of $\mathcal{L}_{\mathsf{Q}}$ (e.g., S may be given as the foundation of mathematics) and if we happen to be able to define the truth of $\mathcal{L}_{\mathsf{Q}}$, then there would be nothing to complain and our concern would shift to the question whether S is indeed acceptable for that purpose independently of truth definitions. However, we are currently considering expanding Q to some S specially for the sake of the truth definition of $\mathcal{L}_{\mathsf{Q}}$. Not to mention Ockham's razor, since we are now interested only in the truth of $\mathcal{L}_{\mathsf{Q}}$ and its analysis, we should focus on truth simpliciter and keep addition and expansion as small as possible. This point poses a difficulty for both (ii°) and (iii°), but it particularly concerns the latter; for, that we are allowed to choose arbitrary such system means that we are allowed to add arbitrary strong assumptions to Q.

Finally, it is to be noted that the expansion of Q may cause a change of the sense of $\mathcal{L}_{\mathsf{Q}}$-terms. As we have seen, the sense of each $\mathcal{L}_{\mathsf{Q}}$-term must be initially fixed by virtue of the background system Q. In order for us to give a truth definition to $\mathcal{L}_{\mathsf{Q}}$, the approach currently considered suggests us to expand $\mathcal{L}_{\mathsf{Q}}$ to a certain language $\mathcal{L}$, adopt some sufficiently rich system S over $\mathcal{L}$

and then define the truth of $\mathcal{L}_{\mathsf{Q}}$ in S. For this purpose, the language $\mathcal{L}_{\mathsf{Q}}$ with the fixed meaning provided by virtue of the initial system Q must be translatable into $\mathcal{L}$. The translation of $\mathcal{L}_{\mathsf{Q}}$ in $\mathcal{L}$ need not coincide with $\mathcal{L}_{\mathsf{Q}}$ as a sublanguage of $\mathcal{L}$ (though they are syntactically identical), but it would not be clear whether there is such a translation if the expansion of Q to S brings a drastic change of the sense of $\mathcal{L}_{\mathsf{Q}}$-terms. In particular, the options (iii°) is implausible from this point of view, since a drastic strengthening of system is very likely to cause a drastic change of meanings.

The approach (i°) thus seems to be most plausible among the above three. However, what one is doing under (i°) is essentially the same as what one is pursuing under the axiomatic approach, which is my own choice and will be explained in detail in the next section. Under the approach (i°), one begins with selecting some principles of truth and then she examines which system can define a predicate satisfying the selected principles. However, the first step is exactly what the axiomatic approach does, and the second step simply amounts to examining whether a certain axiomatic system of truth can be suitably translated (or embedded) in systems in question. For example, the aforementioned fact that NBG defines the predicate satisfying the T-schema for $\mathcal{L}_{\in}$ simply means that the system TB (see §2.2) is syntactically embeddable in NBG. Hence, the axiomatic approach methodologically precedes the approach (i°). In addition, since axiomatic theories of truth are immune from the aforementioned problem concerning self-applicability of truth, it is much more versatile than the definitional approach of the type in question.

1.2 Semantic Approach vs Axiomatic Approach

Now, let us turn to the non-definitional (or non-reductionist) approach. As the attempt to define truth is abandoned in the non-definitional approach, truth must be conceived as a primitive notion of our language which is not further reduced to other notions. This doesn't mean that there is nothing left for the conceptual analysis of truth. We are still interested in the characteristics of

this notion, the role it plays in our linguistic or mathematical activities, its consequences for other notions or other subjects of philosophy or mathematics, and so forth. A non-definitional theory of truth usually amounts to providing a certain formal 'theory' or 'structure' to a language which is already equipped with its own truth predicate (or possibly more than one truth predicates).

Since we have left Tarski's original framework, some terminologies of Tarski should be redefined for the sake of our current framework, i.e., the non-definitional theories of truth. In particular, we need to make it clear what Tarski's notions of meta-language and object-language mean in the context of the non-definitional approach. Originally, Tarski defined them as follows:

> [The object language] is the language which is 'talked about' and which is the subject matter of the whole discussion; the definition of truth which we are seeking applies to the sentences of this language. [The meta-language] is the language in which we 'talk about' the [object-language], and in terms of which we wish, in particular, to construct the definition of truth for the first language. [80, p.125]

In the non-definitional approach, we try to give a certain theory or mathematical structure to a language $\mathcal{L}$ already equipped with some predicate(s) which are intended to be the truth predicate(s) of $\mathcal{L}$. Hence, though we no longer aim at defining the truth of $\mathcal{L}$, it is this $\mathcal{L}$ that is 'the language which is talked about and which is the subject matter of the whole discussion', and I call this $\mathcal{L}$ the object-language in the non-definitional theories of truth on the one hand. On the other hand, the meta-language is the language $\mathcal{L}_{\mathsf{Q}}$ in which the non-definitional theory in question of truth of the object-language $\mathcal{L}$ is given and developed by virtue of a certain background system Q over $\mathcal{L}_{\mathsf{Q}}$. It is to be emphasized that, in order to develop a non-definitional theory of truth, we need to work with some background system Q and I call it the metasystem as in the case of the definitional approach. For example, in the case of Kripke-Feferman theory KF over Peano arithmetic PA ([17]), the object language is $\mathcal{L}_T$, i.e., the language of arithmetic plus a unary truth predicate T, and the metasystem is any system (e.g., $I\Sigma_1$) rich enough to represent the syntax and derivations of primitive recursive systems therein; for another example, in the case of the Kripke's fixed-point semantics over arithmetic ([50]), the object language is the same as the KF case and the

metasystem is a sufficiently rich set theory in which Kripke's construction is possible.

There are various different kinds of 'formal theories of truth'. According to Sheard [76], they are divided into two major categories: i.e., the one is called *semantical* theories of truth and the other is *axiomatic* theories of truth. Sheard doesn't give a detailed explanation of what they are; but the section of 'Axiomatic approaches' in [76] (put next to the section of 'Semantical approaches') start with the following passage: '[a]n alternative way of thinking about self-referential truth is to pose the problem as one of creating not models but consistent theories ([76, p.1044])'. Hence, Sheard seems to consider a semantical theory of truth to amount to a construction of a model-theoretic structure of the object-language in question, while an axiomatic theory aims at giving a consistent set of sentences of the object language which contains some desired principles of truths; therefore, we may well call the former 'a model-theoretic approach' alternatively. Still the distinction between them is no less clear than that between semantics and syntax.[6] Nevertheless, fortunately, as far as the so far presented formal theories of truth are concerned, we can distinguish them in most cases; for instance, Kripke's theory is to be classified as semantical on the one hand, and Kripke-Feferman's theory KF is axiomatic on the other hand.

The origin of the axiomatic approach dates back to Tarski. Faced with the thesis (Tars), Tarski himself wrote:

> [I]t would be incorrect to infer the impossibility of operating consistently and in agreement with intuition with semantical concepts and especially with the concept of truth. But since one of the possible ways of constructing the scientific foundations of semantics is closed we must look for other methods. The idea naturally suggests itself of setting up semantics as a special deductive science with a system of morphology as its logical substructure. For this purpose it would be necessary to introduce into morphology a given

[6] First, it is not definitely clear whether giving a 'consistent' set of sentences over a non-recursively axiomatizable systems. For example, the logic of the system PALTr of Hájek et.al. [28] is infinite-valued Lukasiewicz Logic L$\forall$ which is known to be non-effective (i.e., the set of all 1-tautology is not recursive). However, the motivation of Hájek et.al. is apparently axiomatic; they consider adding some truth-theoretic principles to a given set of arithmetical sentences and show the consistency of the resulting set of sentences with respect to L$\forall$. Second, there is a borderline case between syntax and semantics such as Lindenbaum algebra. Third, since Sheard classifies Barwise and Etchemendy's theories of truth (both Russellian and Austinian ones) as semantical theories, it seems that he does not restrict the sense of the word 'model' to the ordinary model-theoretic sense (with a set domain) and he includes theories which are committed to proper class size structures, such as Barwise and Etchemendy's and Tarski's own, into the category of 'semantical theories of truth'.

> semantical notion as an undefined concept and to establish its fundamental properties by means of axioms. [82, p.255]
>
> [E]ven respect to formalized languages of infinite order, the consistent and correct use of the concept of truth is rendered possibly by including this concept in the systems of primitive concepts of the metalanguage and determining its fundamental properties by means of the axiomatic method (the question whether the theory of truth established in this way contains no contradiction remains for the present undecided). [82, p.266]

From the context in which the second passage is put, it is clear that Tarski mentioned 'formalized languages of infinite order' as an example of the languages for which a formally correct and materially adequate definition of true sentences *cannot* be constructed in the metalanguage; therefore, Tarski himself suggests the axiomatic approach as a remedy for the difficulty with which we are currently concerned. Thus, it is the axiomatic approach that Tarski himself suggested as a natural alternative to the definitional approach. The cited passage also explains what the axiomatic approach is. However, it is not clear whether Tarski endorsed the axiomatic approach, and Tarski himself did not explore this direction further. He only gave some negative comments to one specific axiomatic system TB of truth and closed his debates on the axiomatic approach; the system TB will be introduced in the next chapter and I will come back later to this issue of Tarski's standpoint on the axiomatic approach in the next chapter. I myself prefer the axiomatic approach to the semantical approach and adopt it as the main topic of the present thesis. In the rest of present section, I will argue against the semantical approach.

In contrast to the axiomatic approach –although it is much more popular and dominant among philosophers working on truth today– I am not sure of the origin of the semantical approach. Roughly speaking, the semantical approach starts with a fixed semantical structure of the base language without the truth predicate; then it goes to add semantics of the truth predicate by giving its 'interpretation' or 'extension' over the given structure. In most typical cases such as Kripke's theory [50] and Gupta and Belnap's revision theory [26], the 'semantical structure of the base language' means the ordinary classical model-theoretic structure with a certain set domain. One advantage of this approach is that it is immune from the aforementioned problem on possible

changes of meanings in expanding systems; for, we can naturally assume that the meaning of each term is already fixed in the given *semantics* or model-theoretic structure and it will not be affected by the process of giving the interpretation of the truth predicate. Another advantage is that a semantical theory often gives us more information of the concept of truth in question and make it more perspicuous. However, ultimately, it seems crucially problematic, particularly from the foundational point of view, to require that our metalanguage is given such a model-theoretic or semantical structure from the outset; in other words, I think such an approach is inappropriate for the analysis of the notion of truth of the language which we are indeed using or speaking.

Vänäänen [85] gives a suggestive and useful discussion in this respect. In any context of formal logic, we first fix a 'top-level' system and start our mathematical inquiries from this 'top-level'. Reflecting this fact, Vänäänen points out that:

> Formalization of mathematics involves defining a formal language with some intended meaning. Let us call this language urlogic. The idea is that urlogic is the most primitive formal language we use to study the process of doing mathematics. [85, p.501]
>
> We extracted urlogic as a formalization of the act of doing mathematics. The semantics of urlogic is totally informal. [85, p.501]

The urlogic is the 'top-level' formal counterpart of our informal mathematics, and we do not ascend above urlogic. Let T denote this 'top-level' system. Then, Vänäänen proposes the following criteria for urlogic ([85, p.511]):

- Sentences of urlogic are finite strings of symbols. That a string of symbol is a sentence of urlogic, is a non-mathematical judgement.
- Some sentences are accepted as axioms. That a sentence is an axiom is a non-mathematical judgement.
- Derivation are made from axioms. The derivations obey certain rules of proof. That a derivation obeys the rules of proof, is a non-mathematical judgement.
- Derived sentences can be asserted as facts.

According to these criteria, for example, infinitary logic and full second order logic cannot be urlogic, since they need highly mathematical devices for their formulation; infinitary logic violates

the first condition, and full second order logic violates the second and third. Doing model theory *within* an initially accepted urlogic such as ZF is of course acceptable, but it is very likely to violate these criteria to assume that the metasystem is already equipped with a model-theoretic semantics and the meaning of its language is provided by virtue of it. I will not get into these issues on how metamathematics or the top-level metasystem should be formalized, but I myself strongly agree with Vänäänen's argument.

One possible defense of the semantical approach might be this: in a semantical theory of truth, one is constructing a model-theoretic structure $\mathfrak{A}$ of the language $\mathcal{L}$ *within* a given metasystem which appropriately satisfies Vänäänen's criteria so that the interpreted language $\mathcal{L}$ by $\mathfrak{A}$ has the same meaning of the meta-language. First of all, this argument poses a restriction to the variety of the metasystem: i.e., the metasystem has to be a system which can define and operate on model-theoretic structures. Second, it is not clear at all how they can establish this preservation of meanings between the meta-language and the interpreted language. It seems that this preservation at least requires that the structure $\mathfrak{A}$ should be a model of the meta-system, say, ZF. However, in the ordinary settings, this is impossible due to Gödel's incompleteness and completeness theorems.

Another possible counterargument on behalf of the semantical approach might be this: although many semantical theories of truth operate on model-theoretic structure (as a set-theoretic object), they in fact aim at explaining the procedure of giving semantics to the truth predicate within a suitably expanded metasystem. For example, suppose the initial metasystem is ZF; Then, by expanding ZF to Π^1_1-CA (i.e., NBG plus the comprehension axiom schema for Π^1_1 formulae), we can carry out essentially the same construction of Kripke's interpretation of the truth predicate; we can define the least fixed-point of each elementary positive operator and thereby interpret the Kripkean (self-applicable) truth predicate of $\mathcal{L}_\in$ by means of one such fixed-point. However, this interpretation of what semantical theories of truth do is too charitable and, at any rate, the same problem as I have raised against the definitional approach at the end of the last paragraph applies

to this approach as well.

The arguments against the semantical approach I have raised so far are probably not decisive for rejecting the semantical approach. However, Vänäänen's argument, at least, begs semantical (or model-theoretic) theorists of truth for further justifications of their methodology; we pass the turn back to semantical theorists of truth and it is now their turn to defend their approach.

Now we have chosen the non-definitional axiomatic approach. Let $\mathcal{L}$ be a language equipped with its truth predicate $\mathcal{L}$. We need to block paradoxes and avoid contradiction. Feferman [15] pointed out that, in general, the possible solutions to the paradoxes consist in restrictions of (1°) language, (2°) logic or (3°) basic principles. It is slightly vague what he means by the solution type (1°). Feferman classified Tarski's theory of truth as the solution under (1°). Thus, according to my reading, the option (1°) amounts to restrict the formal study of truth to the language which doesn't contain the truth predicate for itself (it might possibly allowed to contain the truth predicate of a proper sublanguage). As I have argued, Tarski's theory already gives us a fairly satisfactory solution under this category (1°), but we are now searching for another type of theory because of the aforementioned disadvantages of his theory. Hence, we will set aside the option (1°) for the discussion of the present thesis. Then, the remained options are (2°) and (3°). I myself would choose (3°) and will give some arguments against (2°) in the next section.

1.3 Revision of Logic

In the present section, I will consider the option (2°) of revising logic. As I have declared, the objective of the present section is to provide some arguments against this approach. Briefly speaking, my reason for preferring (3°) and avoiding (2°) is: (a) in order to keep the basic principles of our most naïve conception of truth, we apparently need to make too drastic and probably intolerable revisions on our logic; (b) logic is the most fundamental in our reasoning in every subject matter

and thus revising our logic merely for the sake of blocking paradoxes related to the notion of truth is too much; it is as if Einstein had revised the logic for accommodating the incoherency of Newtonian theory of gravity and special theory of relativity rather than to provided general theory of relativity. The problem caused by the Liar paradox and its variants is more naturally conceived as a problem of our way of conceiving truth rather than an intrinsic problem of our logic.

Of course, it is difficult to give a decisive argument for rejecting any approach under (2°); above all, we haven't yet reached the unanimous conclusion of which logic is 'the' logic. Instead, I will illustrate some of the costs we would necessarily have to pay in revising logic, and try to argue that the costs are too high. In the present section, I will present two fairly general paradoxes applicable to a large range of formal logics; thus, they mainly concern the point (a).

First of all, I pose the following assumptions for the following arguments:

(Pred) Truth is expressed by a predicate, i.e., a linguistic device which is applied to singular terms.

(Synt) A reasonably rich theory of syntax is available to any system of truth.

These assumptions look fairly reasonable and broadly accepted among truth theorists. Leitgeb [53, p.277] claims that 'there is almost unanimous agreement' for (Pred); for, we require truth to be applicable to quantified variables so that we can thereby properly express such sentences as 'There is at least one axiom of ZF which is true', and 'For all sentences, its negation is true if and only if it is not true'. I agree with his arguments and take the assumption (Pred) granted. Next, the assumption (Synt) partially concerns the problem of the truth-bearer. Although I am opt for syntactical objects, i.e., more specifically, declarative sentence types, it seems that we haven't yet reached a unanimous conclusion about what the bearers of truth are. However, it is almost beyond doubt that, as long as truth is conceived as a predicate, it is of course to be applied to its bearers and thus any system of truth must be equipped with a reasonably rich theory about the truth-bearers. Therefore, if one prefers another kind of truth-bearers, she may simply replace the

assumption (Synt) by a corresponding assumption on a theory of her choice of truth-bearers. Even if she does so, I do not think that her choice would bring any serious difference to my arguments in what follows. What we will need for the subsequent arguments is a certain structural features of truth-bearers which, I would like to claim, will be reasonably assumed for various other truth-bearers without loss of generality; my stance here coincides with the one Halbach [37] takes; I refer the reader to his book [37, pp.17–20] for more careful and detailed discussions.

Under the type of solution (2°), the language contains its own truth predicate and the basic principles of our naïve conception of truth are tried to be preserved as much as possible. In particular, the principal aim of this approach is to provide a formal system of truth in which we have the full T-schema, i.e.,

$$Tr(\ulcorner\sigma\urcorner) \quad \text{if and only if} \quad \sigma,$$

for arbitrary sentences σ whether or not they may contain the truth predicate Tr itself. We first focus on this requirement of the full T-schema.

It must be determined how the word 'if and only if' should be formalized. First, we need to determine whether it is expressed at the meta-level or object-level.[7] Let S be any system of truth and let $\mathcal{L}$ be its language. Then, the next two different formulations of the T-schema can be considered:

$$\mathsf{S} \vdash \sigma \text{ if and only if } \mathsf{S} \vdash Tr(\ulcorner\sigma\urcorner), \text{ for any sentence } \sigma. \tag{1.2}$$

$$\mathsf{S} \vdash \sigma \leftrightarrow Tr(\ulcorner\sigma\urcorner), \text{ for any sentence } \sigma. \tag{1.3}$$

These two are sometimes mixed up and called a single word 'T-schema', but we have to clearly distinguish them; for example, when we take $\mathsf{S} = \mathsf{FS}$ (the definition will be given in the next

[7] We notice that, in Tarski's theory, this distinction makes no difference, simply because the two formulations are equivalent.

chapter), the former holds while the latter fails. Let us call the former type of equivalence *external T-schema*; i.e., the T-schema expressed by a certain relationship (between sentences σ and $T\ulcorner\sigma\urcorner$) *outside* of the object-language (and within the meta-language). On the other hand, we call the latter type of the equivalence *internal T-schema*; i.e., the T-schema expressed by a certain relationship within the object-language.

Internal formulations of the T-schema has some advantages over external ones. One defect of external formulations consist in the fact that each instance of the T-schema is not a statement of the object-language $\mathcal{L}_\mathsf{S}$; it is a statement of the meta-language. Thus, the speakers of $\mathcal{L}_\mathsf{S}$ cannot assert it at all. This situation brings a difficulty particularly to Horwich's minimalist standpoint for instance. Horwich writes:

> [T]he deflationist maintains that, since our commitment to these schemata [T-schemata] accounts for everything we do with the truth predicate, we can suppose that they implicitly define it. Our brute acceptance of their instances constitutes our grasp of the notion of truth. [39, p.240]

Thus, the minimalist requirement for the truth predicate is not be fulfilled solely by an external formulation of the T-schema from the viewpoint of the speakers of $\mathcal{L}_\mathsf{S}$. Hence, an internal formulation seems preferable to an external one particularly from the minimalist standpoint. It seems that this point is disadvantageous not only for minimalists but also for many other standpoints in general. If each instance of the T-schema is not asserted within the object-language, the T-schema is not the characterization (or, say, 'implicit definition' in terms of Horwich) of truth for the speakers of the object language $\mathcal{L}_\mathsf{S}$ and it is observed only by the speakers of the meta-language that Tr is the truth predicate of $\mathcal{L}_\mathsf{S}$.

Another defect of external formulations specially concerns the currently questioned approach (2°). Recall that the approach (2°) suggests to revise the logic of the object-language in order to thereby accommodate the paradoxes and preserve the full T-schema. In such an approach, the metalogic (i.e., the logic governing the metasystem) and the logic of the object-language may be

different. This gap might be still acceptable. In fact, even truth-theorists endorsing (2°) usually uses classical mathematics such as classical set theory or classical arithmetic at the meta-level in constructing their theory of truth. The resort to classical mathematics at the meta-level could possibly be justified by arguing that they only revise the logic which concerns truth and its relevant notions and keep the ordinary classical mathematics as a separate subject from their discourse about truth. However, if one adopts an external formulation of the T-schema, the biconditional 'if and only if' belongs to the meta-language and follows the law of the metalogic, say, classical logic. Then, this results in a theory of truth which revises the entire logic of the object-language but the target principle of truth follows classical logic; this is a fairly strange consequence and the aforementioned justification for the gap between the meta- and object-logics seems no longer in effect.

Next, we need to determine how the word 'if and only if' is formally expressed. In many cases, we work with systems with one and only 'default' (or 'canonical') logical conditional (and biconditional) and take it for granted that 'if and only if' in the T-schema is expressed in terms of it. However, there may be other formulations and indeed there are already such proposals. For instance, Feferman [15] proposed a system in which 'ϕ iff $Tr(\ulcorner\phi\urcorner)$' is expressed by a formula $\phi \equiv Tr(\ulcorner\phi\urcorner)$ with an newly introduced special connective '$\equiv$', which can be interpreted as 'ϕ and $Tr(\ulcorner\phi\urcorner)$ have the same truth value (in Strong Kleene Kripkean fixed-point model)'[8]; this system is called Aczel-Feferman system AF.

1.3.1 Moh Shaw-Kwei's Paradox

In the present subsection, I introduce the so-called Moh Shaw-Kwei's paradox which was found by Moh Shaw-Kwei [75]. The following formulation of mine is a slightly generalized version of the

[8]This interpretation is not due to Feferman himself but to Sheard [76]. Feferman himself seems to intend to characterize the connective '$\equiv$' in a purely syntactical manner.

original one.

Although we are now considering the axiomatic approach, I will take a more general setting so as to inclusively treat both axiomatic and semantical theories of truth in what follows. In spite of the division of axiomatic and semantical theories, they have one common feature: namely, they are supposed to specify a certain special class of 'designated' sentences among the object-language. When a theory is semantical, such a class can be defined as a class of 'satisfied' sentences in a given semantical structure; on the other hand, when a theory is axiomatic, the class of 'provable' sentences corresponds to it. This view seems to be shared among those who are working on formal theories of truth.[9] Throughout the present subsection, the word 'system' will denote any mathematical structure for a certain language that decides which sentences are designated, regardless of whether it is semantical or axiomatic. This definition of 'system' here is extremely general. Set-theoretically, a system Q is simply represented by a triple $\langle \tau, \mathcal{L}_{\mathsf{Q}}, \models_{\mathsf{Q}} \rangle$, where τ is a first-order vocabulary containing a truth predicate Tr, $\mathcal{L}_{\mathsf{Q}}$ is the language of Q including the first-order language formed from τ (i.e., $\mathcal{L}_{\omega\omega}[\tau]$ in abstract model-theoretical terminology), and $\models_Q$ is a subset of $\mathcal{L}_Q$ representing the set of designated $\mathcal{L}_{\mathsf{Q}}$-sentences. For instance, each Kripkean fixed-point structure $\mathfrak{A}^*$ for a language $\mathcal{L}$ is a system in this sense; we say that a sentence ϕ of $\mathcal{L}$ is designated in $\mathfrak{A}^*$ when $\mathfrak{A}^* \models \phi$. For other examples, in any first-order axiomatic system T, ϕ is designated when $\mathsf{T} \vdash \phi$; in any theory based on n-valued Lukasiewicz logic, ϕ is designated when the value assigned by the theory to ϕ is 1.

Definition 1.3.1. Let Q be any system of truth and let I and C be definable binary sentential connectives of $\mathcal{L}_{\mathsf{Q}}$. I is said to be *detachable* in Q, if the following holds:

$$\text{if } I(\phi, \psi) \text{ and } \phi \text{ are designated in } \mathsf{Q}, \ \psi \text{ is designated in } \mathsf{Q}.$$

[9] For example, Sheard endorses this view. According to him, 'formal representation' of truth refers to 'any formal framework, either semantic or syntactic, which results in the validation of a set of sentences, including some containing a truth predicate ([61, p.169])'.

I is said to be *transitive* in Q, if the following holds:

if $I(\phi,\psi)$ and $I(\psi,\theta)$ are designated in Q, $I(\phi,\theta)$ is designated in Q.

Sentences ϕ and ψ are said to be congruent, when

if θ is designated, the formula θ' is designated,

for any sentence θ, where θ' is obtained from θ by replacing ϕ (or ψ, resp.) in arbitrary places by ψ (or ϕ, resp.). C is called a congruence in Q, if ϕ and ψ are congruent whenever $C(\phi,\psi)$ is designated.

Remark 2. In Priest's system T_0 ([67]), its (default) conditional is neither detachable nor transitive and its (default) biconditional is not a congruence. In Aczel-Feferman AF, the newly introduced connective '$\equiv$' is not a congruence; in addition, even when $\phi \equiv \psi$ holds, ϕ (ψ, resp.) does not necessarily implies ψ (ϕ, resp.); see [15, p.98].

Definition 1.3.2. Let I be a definable binary sentential connective in a system Q. For formulae ϕ and ψ, $I^n(\phi,\psi)$ is defined recursively by

$$I^1(\phi,\psi) \text{ is } I(\phi,\psi)$$
$$I^{k+1}(\phi,\psi) \text{ is } I(\phi, I^k(\phi,\psi)).$$

Definition 1.3.3. Let I be a definable binary connective in a system Q. We say that the rule of absorption of order n, denoted $(\mathrm{A})_n$, holds (or admissible) with respect to I in Q, if the following holds:

If $I^{n+1}(\phi,\psi)$ is designated in Q, then $I^n(\phi,\psi)$ is designated in Q.

Remark 3. $(\mathrm{A})_1$ amounts to the so-called Contraction Rule and it holds with respect to the canonical conditional in many familiar logics. For another example, $(\mathrm{A})_{n-1}$ holds in n-valued Lukasiewicz logic.

Definition 1.3.4. A system Q is said to be *non-explosive*, if there exists at least one sentence $\bot \in \mathcal{L}_{\mathsf{Q}}$ which is not designated in Q. We call such a sentence $\bot$ a *bottom* of Q.

It is fairly reasonable to assume that every system we consider is non-explosive.[10] I will assume the following throughout the following:

(Bot) Every system has a bottom.

Definition 1.3.5. Let Tr be a predicate of $\mathcal{L}_{\mathsf{Q}}$. This predicate Tr is intended to be the truth predicate of $\mathcal{L}_{\mathsf{Q}}$.

(1) Let I be a binary connective. We say that Q admits the internal T-schema w.r.t. I, iff $I\big(\phi, Tr(\ulcorner\phi\urcorner)\big)$ and $I\big(Tr(\ulcorner\phi\urcorner), \phi\big)$ are designated in Q for all $\phi \in \mathcal{L}_{\mathsf{Q}}$.

(2) We say Q admits the congruent external T-schema, iff ϕ and $Tr(\ulcorner\phi\urcorner)$ are congruent for any $\phi \in \mathcal{L}_{\mathsf{Q}}$.[11]

Definition 1.3.6. Let Q be a theory and let P a unary predicate of $\mathcal{L}_{\mathsf{Q}}$.

Let I be a binary connective of $\mathcal{L}_{\mathsf{Q}}$. We say that P admits internal self-reference with respect to I, iff there is a sentence $\lambda \in \mathcal{L}_{\mathsf{Q}}$ such that

$$I\big(\lambda, P(\ulcorner\lambda\urcorner)\big) \text{ and } I\big(P(\ulcorner\lambda\urcorner), \lambda\big) \text{ are designated.} \tag{1.4}$$

We say that P admits congruent external self-reference, iff there is a sentence λ such that λ and $P(\ulcorner\lambda\urcorner)$ are congruent in Q.

[10] Even a system in paraconsistent logic is supposed to have a bottom. Paraconsistent logic is a logic which renounces the law of non-contradiction but at the same time it does not allow systems to be trivial or explosive in the sense that all sentences are designated.

[11] Note that the congruence between ϕ and $Tr(\ulcorner\phi\urcorner)$ is a relation at the meta level.

Remark 4. Here, one specific formulation of self-reference is implicitly assumed: that is, a self-referential sentence σ w.r.t. a predicate P is such a sentence that σ iff $P(\ulcorner\sigma\urcorner)$. There are several reasons to prefer this formulation of self-reference: as Gödel showed, if a theory is provided with its structural descriptives, this form of self-reference is expected to be derived; also, in most cases, either congruent external or internal self-reference of this form follows from another formulation of self-reference and the T-schema. Of course, there are other formulations of self-reference. One natural alternative is to formulate a self-referential sentence σ w.r.t. P as a sentence $\ulcorner P(\ulcorner\sigma\urcorner)\urcorner = \ulcorner\sigma\urcorner$. Another is that of the revision theory [26] via "definitional equivalence". The third is the so-called situation theoretical interpretation of self-reference.[12]

Theorem 1.3.7. Let Q be a system with a bottom $\bot$, I a binary connective in Q and Tr a unary predicate of $\mathcal{L}_{\mathsf{Q}}$. Suppose that $(\mathrm{A})_n$ holds in Q for some $n \in \mathbb{N}$.

(i) If I is transitive and detachable and the predicate $I^n(Tr(x), \bot)$ admits internal self-reference with respect to I, then Q does not admit the internal T-schema w.r.t. I.

(ii) If I is detachable and the predicate $I^n(Tr(x), \bot)$ admits internal self-reference w.r.t. I, then Q does not admit the congruent external T-schema.

(iii) If I is detachable and Q admits the internal T-schema w.r.t. I, then the predicate $I^n(Tr(x), \bot)$ does not admit congruent external self-reference.

(iv) If I is detachable and $I(\phi, \phi)$ holds for all $\phi \in \mathcal{L}_{\mathsf{Q}}$ and if the predicate $I^n(Tr(x), \bot)$ admits congruent external self-reference, then Tr does not admit the congruent external T-schema.

Proof. **(i)** Suppose Tr admits the internal T-schema w.r.t. I. Since $I^n(Tr(x), \bot)$ admits internal

[12] A typical example of the situation theoretical interpretation can be found in Barwise and Moss's theory [1, Ch.13]. The intuition behind the Barwise-Moss theory is that the utterance of a paradoxical sentence such as the Liar brings a shift of situation or context, and, as a consequence of this shift, a genuine self-reference fails to be made in such a sentence.

self-reference w.r.t. I, there exists a sentence λ such that

$$I\big(I^n(Tr(\ulcorner\lambda\urcorner), \bot),\, \lambda\big) \quad \text{and} \quad I\big(\lambda,\, I^n(Tr(\ulcorner\lambda\urcorner), \bot)\big).$$

By the internal T-schema w.r.t. I, we have

$$I\big(Tr(\ulcorner\lambda\urcorner),\, \lambda\big) \quad \text{and} \quad I\big(\lambda,\, Tr(\ulcorner\lambda\urcorner)\big).$$

Thus, by the transitivity of I we obtains

$$I\big(Tr(\ulcorner\lambda\urcorner),\, I^n(Tr(\ulcorner\lambda\urcorner), \bot)\big) \quad \text{and} \quad I\big(I^n(Tr(\ulcorner\lambda\urcorner), \bot),\, Tr(\ulcorner\lambda\urcorner)\big).$$

From the first, using $(\mathrm{A})_n$, we have $I^n(Tr(\ulcorner\lambda\urcorner), \bot)$. Then, $Tr(\ulcorner\lambda\urcorner)$ is obtained from the second by the detachability of I. Finally, by applying $Tr(\ulcorner\lambda\urcorner)$ n-times to $I^n(Tr(\ulcorner\lambda\urcorner), \bot)$, we obtain $\bot$ in Q; a contradiction.

(ii) Suppose Q admits the congruent external T-schema. Since $I^n(Tr(x), \bot)$ admits internal self-reference w.r.t. I, there exists a sentence λ such that

$$I\big(I^n(Tr(\ulcorner\lambda\urcorner), \bot),\, \lambda\big) \quad \text{and} \quad I\big(\lambda,\, I^n(Tr(\ulcorner\lambda\urcorner), \bot)\big).$$

By the congruence of $Tr(\ulcorner\lambda\urcorner)$ and λ, we obtain that

$$I\big(I^n(Tr(\ulcorner\lambda\urcorner), \bot),\, Tr(\ulcorner\lambda\urcorner)\big) \quad \text{and} \quad I\big(Tr(\ulcorner\lambda\urcorner), I^n(Tr(\ulcorner\lambda\urcorner), \bot)\big)$$

The rest of proof is the same as **(i)**.

(iii) Similar to **(ii)**.

(iv) By the law of identity and congruent external self-reference, we have

$$I\big(Tr(\ulcorner\lambda\urcorner), \lambda\big) \quad \text{and} \quad I\big(\lambda, Tr(\ulcorner\lambda\urcorner)\big).$$

Then, the congruent external T-schema implies that

$$I\big(Tr(\ulcorner\lambda\urcorner), I^n(Tr(\ulcorner\lambda\urcorner), \bot)\big) \quad \text{and} \quad I\big(I^n(Tr(\ulcorner\lambda\urcorner), \bot), Tr(\ulcorner\lambda\urcorner)\big).$$

The rest of proof is the same as **(i)**. □

Several well-known logics fall among the victims of Theorem 1.3.7: more precisely, their default conditionals '$\rightarrow$' are detachable and transitive and they admit $(\mathrm{A})_n$ with respect to '$\rightarrow$'. In particular, among such logics are the following:

- Classical logic and intuitionistic logic;
- Finite-valued Lukasiewicz logic and Post logic[13];
- Weak and strong Kleene Logics and Feferman Logic[14];
- Infinite-valued and finite-valued Gödel Logic[15];
- Any substructural logic with Associativity and Weak Contraction, wherever I is a left-to-right conditional.[16]

Consequently, whenever a system S of truth follows a logic among them and the T-schema and self-reference are formulated internally with respect to its default conditional, (i)-(iii) of Theorem

[13]For the definition and explanation of Post Logic, see [55].
[14]We will explain Feferman logic later in the next chapter.
[15]For the definition and explanation of Gödel Logic, see [27] or [55]. As its name indicates, finite-valued Gödel Logic was introduced by Gödel and used to prove that any finite-valued logic cannot characterize intuitionistic logic. Infinite-valued version is sometimes called Gödel-Dummett Logic.
[16]I follows the Terminology of [72] here. In such a substructural logic, we have, for instance, that if $X \vdash A \rightarrow (A \rightarrow B)$ then $X \vdash (A \rightarrow B)$.

1.3.7 apply to S. Moreover, since each logic above, except for Weak and Strong Kleene Logics and Post Logic, admits $\phi \to \phi$, (iv) of Theorem 1.3.7 applies to any such system.

The possible solutions for this MSK paradox are:

(a) To express the T-schema in the way that 'if and only if' is neither congruent nor transitive and detachable;

(b) To deny (A_n) for any n with respect to I;

(c) To restrict self-reference.

I will not consider (c) here. In what follows, I will examine the solutions (a) and (b).

Solutions of the type (a) are problematic from the viewpoint of the usage and utility of truth predicate. It is probably reasonable to say that we extract the T-schema as a necessary principle of truth from our ordinary usage and utility of the notion of truth. Hence, if we call a formal framework S a theory of the concept of X, S must reflect ordinary usage of the notion of X in particular cases. Reflecting this general requirement for theories of truth, McGee raises the following requirement.

(O) Ordinary Usage Requirement. A successful theory of truth ought to agree with ordinary usage about the applicability of 'true' in a wide range of particular cases. [58, p.159]

Our everyday life usage of the truth predicate, it seems obvious that we use and apply the T-schema with the combination of the transitivity and detachability of the conditional 'if ... then ...'. For example, when we infer 'Ann's claim is true iff $P \neq NP$' from 'Bob conjectures that $P \neq NP$ is true' and 'Ann claims that Bob's conjecture is true', we apparently use the T-schema and the transitivity of 'if ... then ...' (twice) in this inference; if $P \neq NP$ we infer 'Ann's claim is true' by appealing to the detachability of 'if ... then ...'. I dare not bluntly reject the option (a) simply appealing to the thesis (O). As we have mentioned, Feferman's AF takes this strategy; for another example, the default conditional of Priest's theory T_0 is neither detachable nor transitive. However, at any rate, any approach with (a) is required to justify its violation of (O).

Let us turn to the option (b). As far as I know, we already have three systems of truth which satisfy everything we have so far demanded except $(\mathrm{A})_n$: PALTr of Hájek et.al. ([28]), Priest and Sylvan's T_2 ([67]) and Field's theory ([22]). In these systems, the default conditional '$\rightarrow$' is detachable and transitive, the default biconditional '$\leftrightarrow$' is a congruence, $\phi \rightarrow \phi$ holds for any ϕ, and both the internal (and congruent external) T-schema and self-reference are met; but, of course, $(\mathrm{A})_n$ fails for all $n \in \mathbb{N}$. Admittedly, the option (b) is the most straightforward solution to the MSK paradox. However, there remain some issues to be examined.[17]

First, we must ask whether the absorption rule $(\mathrm{A})_n$ is indeed dispensable for theories of truth. Compared to detachability and transitivity, dropping $(\mathrm{A})_n$ looks much less harmful. However, the next proposition indicates that $(\mathrm{A})_n$ bears some close connection to the monotonicity of our reasoning.

Proposition 1.3.8. Let L be a finite-valued logic L (in the standard sense of [55]) and let I be a definable binary sentential connective. We denote the value of a sentence ϕ by $||\phi||$. Suppose L has the following property: for each sentence ϕ and ψ,

(Monotonicity) $||\psi|| \leq ||I(\phi, \psi)||$;

this condition is naturally interpreted as expressing that the reasoning is monotonous in the sense that the more assumptions we have the more consequences we obtain. Then, for all sentences ϕ and ψ, there exists $n \in \mathbb{N}$ such that $||I^{n+1}(\phi, \psi)|| = ||I^n(\phi, \psi)||$.

Notice that the proof of Theorem 1.3.7 indeed only requires some specific sentences ϕ and ψ (i.e., $Tr(\ulcorner\lambda\urcorner)$ and $\bot$) to meet $||I^{n+1}(\phi, \psi)|| = ||I^n(\phi, \psi)||$. Hence, if we adopt a finite-valued logic, Theorem 1.3.7 forces us to somehow restrict the monotonicity of our reasoning unless we adopt infinite-valued logic.[18] As a corollary, since the conditions of the last theorem are met in Field's

[17] Each theory has its particular problems. For instance, PALTr is not recursively axiomatizable and has no standard model. According to Leitgeb [53], it is required as a desideratum for a theory of truth that the theory should allow for standard interpretations. Thus, PALTr is problematic from this viewpoint.

[18] Going beyond finite-valued logic is not as outrageous as one may expect at a first glance. Non-finite valued logic

theory, it follows that the logic of Field's theory (though it is semantical) is not any finite-valued logic.

Second, systems based on such logics might possibly not be of sufficient use. For instance, it is to be examined whether a sufficient part of arithmetic or physics can be described with such logics. Lastly, even if we block the MSK paradox by abandoning the absorption rule, other paradoxes might arise. As a matter of fact, there *is* another paradox, which arises from some additional assumptions but without any need to resort to $(\mathrm{A})_n$; we will see one such paradox in the next section.

1.3.2 Hajek-Paris-Shepherdson Paradox

Tarski raised the full T-schema as the formally correct and material adequate condition to be met by any definition of truth. It is a subtle issue whether he would still maintain this condition even for non-definitional theories of truth; he would probably agree that the derivability of the T-schema is still a desideratum of a non-definitional theory of truth, but he might reject that it is also sufficient. We will come back later to this issue in the next chapter after introducing some axiomatic systems of truth.

At any rate, some further conditions may well be demanded as desiderata of a satisfactory theory of truth. One possible such condition is the following type of 'compositional axiom':

$$Tr\big(\ulcorner \circ(\phi_0, \ldots, \phi_k)\urcorner\big) \quad \text{if and only if} \quad \circ\big(Tr(\ulcorner\phi_0\urcorner), \ldots, Tr(\ulcorner\phi_k\urcorner)\big), \tag{1.5}$$

where $\circ$ is $k+1$-ary logical connective. Again, we are compelled to choose external or internal formulation of 'if and only if'. In what follows, I will only consider the internal formulation and assume that it is expressed by a detachable and transitive sentential connective; for the sake of readability, I will use $\rightarrow$ instead of I. The new desideratum is then formulated as follows:

---

can be effectively axiomatizable of course; for example, it is well-known that intuitionistic logic is not a finite-valued logic.

(Internal Strong Compositionality) For each logical connective $\circ$,

$$\forall x\big[Tr\big(\dot{\circ}(\vec{x})\big) \leftrightarrow \circ\big(Tr(x_0), \ldots, Tr(x_k)\big)\big],$$

where $\dot{\circ}$ is the represetntation of the syntactical operation $\circ(\phi_0, \ldots, \phi_k)$ such that

$$\dot{\circ}(\ulcorner\phi_0\urcorner, \ldots, \ulcorner\phi_k\urcorner) = \ulcorner\circ(\phi_0, \ldots, \phi_k)\urcorner.$$

Let us abbreviate this condition by ISR. I will not get into a substantial discussion on whether ISR is indeed dispensable for a satisfactory theory of truth. However, there are some arguments in favor of ISR. For instance, it is sometimes argued that the T-schema alone is too weak for a theory of truth. Indeed, as we will see in the next chapter, Tarski himself disregards a system TB simply because it only derives the T-schema and lacks sufficient deductive power; for instance, some generalizations of important logical laws such as the excluded middle and the principle of contradiction are not derivable in TB.[19] In order to compensate this defect, as Halbach argues in [35], one might well strengthen the theory by adding ISR. According to Halbach, 'the 'inductive' clauses have been proven to be natural axioms and all generalizations not provable from them seem to be better left undecided by a good theory of truth', where by 'inductive clauses' he means the compositional truths in the above form (1.5); for another example, Field [21] expresses that 'I think it is clear that without such general laws the truth [i.e., the inductive clauses] predicate would not serve its main purpose'. Of course, these arguments are basically concerning classical systems and the T-schema restricted to the base language and thus they might not be directly applicable to our current discussion; but, at any rate, ISR is certainly worth serious consideration.

[19] In general, Halbach [34] showed that any truth-theoretic equivalents of infinite conjunction, i.e., $\forall x(\phi(x) \rightarrow Tr(x))$ where the extension of ϕ is infinite, is not provable in TB. Actually, by slightly modifying his proof, we have a strengthened version of this theorem: i.e., any truth-theoretic equivalents of infinite conjunction is not provable in any system obtained from the base system B by adding a *consistent* T-schema restricted to some fragment of $\mathcal{L}_B \cup \{Tr\}$.

The present subsection will reveal that ISR causes another paradox when combined with the internal T-schema and self-reference (w.r.t. '$\to$'). I call it Hájek-Paris-Shepherdson paradox (HPS paradox, for short), because the prototype of this paradox was presented by Hájek, Paris and Shepherdson [28]. This paradox shows a quite different way of producing a contradiction; in particular, it does not need the absorption rule $(\mathrm{A})_n$.

Compared to the Liar and the MSK paradox, the HPS paradox is slightly complicated. What I will state as the 'HPS paradox' below is just one possible formulation, and there are many other formulations. The following formulation might look ad hoc, but the formulation itself is not important; rather, the background mechanism and reasoning in its yielding a contradiction is important. I will restrict the subsequent arguments to the axiomatic systems of truth, but similar arguments can be applied to semantical theories of truth with some straightforward modifications.

Throughout the following, I will assume that the object (axiomatic) system S (over its language $\mathcal{L}_{\mathsf{S}}$) includes the first-order quantificational language of arithmetic and Primitive Recursive Arithmetic PRA formulated thereover; i.e., all PRA-theorems are designated in S. Hence, a primitive recursive function $F(x,y)$ such that $F(n,m) = \ulcorner \phi \to^n \bot \urcorner$ is contained in S, where m is the code of ϕ and $\theta_0 \to^n \theta_1$ is defined analogously from $I^n(\theta_0, \theta_1)$. I will also assume in the following that the object system admits the universal instantiation (UI) with respect to '$\to$': i.e., $\forall x \phi(x) \to \phi(t)$ for any term t.

Informally, HPS paradox says that: assuming moderate and natural rules about existential quantification and conditional,

- If ISR holds in S, then S is ω-inconsistent (in a modified sense).
- If ISR holds in S and arithmetical induction schema for every formula of $\mathcal{L}_{\mathsf{S}}$ is derivable in S, then S is inconsistent.

Let us see the formal statements of these.

Theorem 1.3.9. If the conditions (i)-(v) below hold in S and ISR holds in S, then S is ω-inconsistent; i.e., there is a (first-order) $\mathcal{L}_{\mathsf{S}}$-formula ϕ such that $\exists x\phi$ is designated in S but $\phi(\bar{n})$ is designated for no $n \in \mathbb{N}$, where $\bar{n}$ is the numeral for n.

(i) If $\phi \to \exists x\psi$ is designated and x does not freely occur in ϕ, then $\exists x(\phi \to \psi)$.

(ii) If $t = s$ is designated, then t and s are interchangeable (w.r.t. designation) in any formula.

(iii) $\phi(t) \to \exists x\phi$ is designated.

(iv) If $\phi \to \psi$ is designated, then so is $\exists x\phi \to \exists x\psi$.

(v) If $\exists x\phi$ is designated and x is not free in ϕ, then ϕ is designated.

Proof. By the internal T-schema, we take a sentence λ such that $\lambda \to \exists x Tr(F(x, \ulcorner\lambda\urcorner))$ and $\exists x Tr(F(x, \ulcorner\lambda\urcorner)) \to \lambda$. Then, we obtain:

(1) $Tr(\ulcorner\lambda\urcorner) \to \exists x Tr(F(x, \ulcorner\lambda\urcorner))$ by the internal T-schema and transitivity;

(2) $\exists x\big[Tr(\ulcorner\lambda\urcorner) \to Tr(F(x, \ulcorner\lambda\urcorner))\big]$ by (1) and (i);

(3) $\big[Tr(\ulcorner\lambda\urcorner) \to Tr(F(x, \ulcorner\lambda\urcorner))\big] \to \big[Tr\big(\ulcorner\lambda\urcorner \dot{\to} F(x, \ulcorner\lambda\urcorner)\big)\big]$ by ISR and (UI);

(4) $\big[Tr(\ulcorner\lambda\urcorner) \to Tr(F(x, \ulcorner\lambda\urcorner))\big] \to \big[Tr(F(x+1, \ulcorner\lambda\urcorner))\big]$ by (3) and (ii);

(5) $Tr(F(x+1, \ulcorner\lambda\urcorner)) \to \exists x Tr(F(x, \ulcorner\lambda\urcorner))$ by (iii);

(6) $\big[Tr(\ulcorner\lambda\urcorner) \to Tr(F(x, \ulcorner\lambda\urcorner))\big] \to \exists x Tr(F(x, \ulcorner\lambda\urcorner))$ by (4) and (5);

(7) $\exists x\big[Tr(\ulcorner\lambda\urcorner) \to Tr(F(x, \ulcorner\lambda\urcorner))\big] \to \exists x \exists x Tr(F(x, \ulcorner\lambda\urcorner))$ by (6) and (iv);

(8) $\exists x \exists x Tr(F(x, \ulcorner\lambda\urcorner))$ by (2) and (7);

(9) $\exists x Tr(F(x, \ulcorner\lambda\urcorner)))$ by (8) and (v);

However, if we have $Tr(\lambda \to^n \bot)$ for some $n \in \mathbb{N}$, since λ follows from (9) by our choice of λ, we could obtain $\bot$ by the detachability of '$\to$'. Thus, by (Bot), $Tr(\lambda \to^n \bot)$ holds for no n. $\square$

Definition 1.3.10. The induction rule IND is defined by

IND**:** For each formula $\phi(x)$, if $\phi(0,\vec{v})$ and $\forall x\big(\phi(x)\to\phi(x+1)\big)$ are designated, then $\forall x\phi(x,\vec{v})$ is designated.

Theorem 1.3.11. Suppose that the conditions (i)-(v) of the last theorem and (vi)-(vii) below are all designated in S. If ISR and IND are also designated in S, then S is inconsistent.

(vi) If $\phi\to(\psi\to\theta)$, then $\psi\to(\phi\to\theta)$.

(vii) If $\phi\to\psi$, then $(\psi\to\theta)\to(\phi\to\theta)$.

Proof. $Tr(\lambda\to^0\bot)\to\bot$, i.e., $Tr(\bot)\to\bot$, immediately follows from the T-schema. Thus, the base step is established.

(1) $Tr\big(\ulcorner\lambda\urcorner\dot{\to}F(t,\ulcorner\lambda\urcorner)\big)\to\big(Tr(\ulcorner\lambda\urcorner)\to Tr(F(t,\ulcorner\lambda\urcorner))\big)$ by ISR;

(2) $Tr(F(t+1,\ulcorner\lambda\urcorner))\to\big[Tr(\ulcorner\lambda\urcorner)\to Tr(F(t,\ulcorner\lambda\urcorner))\big]$ by (1) and (ii);

(3) $Tr(\ulcorner\lambda\urcorner)\to\big[Tr(F(t+1,\ulcorner\lambda\urcorner))\to Tr(F(t,\ulcorner\lambda\urcorner))\big]$ by (2) and (vi);

(4) $Tr(F(t+1,\ulcorner\lambda\urcorner))\to Tr(F(t,\ulcorner\lambda\urcorner))$ by (3), since λ is designated by the last theorem;

(5) $\big[Tr(F(t,\ulcorner\lambda\urcorner))\to\bot\big]\to\big[Tr(F(t+1,\ulcorner\lambda\urcorner))\to\bot\big]$ by (4) and (vii);

Thus, by IND, we obtain

$$\forall x[Tr(F(x,\ulcorner\lambda\urcorner))\to\bot]$$

It follows from (iv) that

$$\exists xTr(\lambda\to^x\bot)\to\exists x\bot$$

However, since we have λ, $\exists xTr(\lambda\to^x\bot)$ holds; a contradiction. □

Note again that, as I have already mentioned, there must be many other sets of conditions for the HPS paradox than (i)-(v) (and (vi)-(vii)). In any case, I think that Theorems 1.3.9 and 1.3.11 together with MSK paradox already well illustrate how very much the revision of logic costs.

1.4 The Classical Axiomatic Approach

I have so far argued that the classical axiomatic approach is more versatile and methodologically superior to the other approaches that we have examined. Hence, as I have declared at the beginning, my choice is the classical axiomatic approach toward the conceptual and logical analyses of the notion of truth. In this approach, we examine various principles of truth in the form of axiom or inference rule over a certain axiomatic deductive system.

We have decided not to revise the logic and keep it classical.[20] Hence, this approach inevitably imposes us a revision of our naïve conception of truth and some basic principles of truth derived from this naïve conception, since they together lead to a contradiction. However, as long as we have decided to abandon our very naïve conception of truth, we have lost general guiding principles for constructing systems of truth. There are many possibilities, conditions and options to be considered. Put in this situation, we need, besides traditional kinds of philosophical speculations, to conduct tedious but down-to-earth 'field work' in systems of truth by examining and comparing them from various points of view; we may carry out experimental constructions of new systems, look at various systems in connection to the other relevant areas of philosophy and mathematics or investigate what kind of and how rich mathematical consequences would be brought to us by each truth-theoretical means. The rest of the present thesis is devoted to this task.

I will propose two new means to compare more conceptual aspects of theories of truth than

[20]I have not given any particular arguments for adopting classical logic, but I have argued that the revision of the logic should be kept as small as possible and we should rather reconsider the basic principles of truth without hastily revising the logic merely for preserving the principles in question. Although some people claim that intuitionistic or other non-classical logics should be adopted as our logic, still most people are working with classical logic and I can see no reason to start by rejecting it.

other so far suggested means. The one is *relative truth definability* discussed in Chapter 3; the other is *comparison by inner theories* in Chapter 4. Before arguing them, in the next chapter, I will introduce the systems of truth to be studied in the present thesis and present basic results of them; most of them are not due to myself but some new results are contained there as well.

It is to be noted that the aforementioned problem on possible changes of meaning in expanding a system may well apply to the axiomatic approach as well. When we formulate an axiomatic truth system over a given base system, it often happens that the resulting system becomes stronger than the base system and brings new non-truth-theoretical consequences (i.e., new theorems of the language of the base system). Some axiomatic truth system over Zermelo's set theory Z may yield the axiom schema of Replacement; some system over ZF may derive the existence of an inaccessible cardinal. If such happens, it may be the case that the system is not about a system of truth of the original base language but rather a system of truth of some other language. For this reason, I have certain sympathy with the *conservativity thesis* which claims that truth systems must be conservative over their base system (for the consequences of the base language), simply because conservative systems are most likely to preserve the meaning of the original base language; this conservativity thesis is largely accepted by deflationists of truth and it is sometimes argued that this thesis is essential to deflationism (cf. [74]). However, it is not clear that non-conservative systems always bring about certain changes on the meaning to some inacceptable degree; also, it is not necessarily assumed that adding the truth predicate to the language should never cause any changes on the meanings of the original languages. I will not go into this issue any more in the present thesis and leave it for another occasion. Instead, I mention another import and significance of the axiomatic study of truth. The above debates on conservativity concerns the question how the notion of truth is to be characterized by means of axioms. One may ask another question: what does it mean to accept a certain system as true? There must be some essential gap between one's simply possessing a truth predicate of a language $\mathcal{L}$ with its proper meaning and

her accepting a mathematical system over $\mathcal{L}$ as true; the latter should be committed to something more and stronger than the former. I will come back to this aspect of the axiomatic study of truth in Chapter 5, where I discuss and study what I call truth progression.

Chapter 2

List of Systems of Truth

In the present chapter, we will introduce various axiomatic systems of truth and survey the basic results about them. Before introducing those systems, we begin with some preliminary definitions.

2.1 Notational Preliminaries

We can talk of the truth of any kind of subject. There are various options for base languages and systems over which we formulate axiomatic truth. In the present thesis, I follow the traditional setting and adopt an arithmetical language and system as the base system. We start by fixing an arithmetical system B as our base system to which every system of truth is obtained by adding truth predicate(s) and its (their) axioms and inference rules.

In what follows, the base system B is assumed to be a first-order classical system of arithmetic. We are principally interested in the case where B is Peano Arithmetic PA, but we adopt a slightly more general setting particularly in order to deal with the relative truth definability among Feferman's schematic reflective closures (see §10) in the same framework as other systems (since we need to add an extra predicate to the base language in formulating schematic reflective closures); there-

fore, after Chapter 3 where we will discuss relative truth definability, we will restrict our discussion to the cases where B is PA in Chapters 4 and 5. The language $\mathcal{L}_{\mathsf{PA}}$ of PA has equality as its only relation symbol and constant $\bar{0}$ for zero as its only constant symbol. For some technical reasons, we also assume that $\mathcal{L}_{\mathsf{PA}}$ contains function symbols for each primitive recursive function and PA contains axioms defining them. Given a language $\mathcal{L}' \supset \mathcal{L}_{\mathsf{PA}}$, *full induction for* $\mathcal{L}'$ is the following schema:

$$\phi(0) \wedge \forall x\big(\phi(x) \to \phi(x+1)\big) \to \forall x \phi(x), \text{ for each } \mathcal{L}'\text{-formula } \phi.$$

Then, our base system B is any extension of PA such that:

(a) Its language $\mathcal{L}_{\mathsf{B}}$ comprises $\mathcal{L}_{\mathsf{PA}}$ and possibly finitely many auxiliary relation symbols.

(b) Its axioms may include a primitive recursive subschema of full induction for $\mathcal{L}_{\mathsf{B}}$.

(c) B may contain at most finitely many axioms other than the specified in (b).

(d) It is arithmetically sound: namely, B is satisfiable by some expansion of the standard model $\mathbb{N}$; in other words, there is a model $\mathfrak{M}$ of B whose $\mathcal{L}_{\mathsf{PA}}$-reduct coincides with $\mathbb{N}$.

We fix such a base system B throughout the present and next chapters and $\mathcal{L}_0$ denotes the language $\mathcal{L}_{\mathsf{B}}$ of B. All the statements in the following are valid regardless of the choice of the base system B unless otherwise stated, subject to the conditions (a)-(d). For a given system Q, $\mathcal{L}_{\mathsf{Q}}$ denotes the language of Q. The language of every system of truth presented in this thesis is an expansion of $\mathcal{L}_0$ with a primitive recursive set of auxiliary predicate symbols. Hence, for any system Q of truth in the present paper, the $\mathcal{L}_{\mathsf{Q}}$-terms are the same as the $\mathcal{L}_{\mathsf{PA}}$-terms.

In formulating a system Q of truth, we have to fix some coding (Gödel numbering) of the language to which the truth predicate is to be applied. First of all, we assume that, for any language $\mathcal{L}$, the logical connectives are '$\neg$' (negation), '$\wedge$' (conjunction) and '$\to$' (conditional) and

the quantifier is '$\forall$' (universal quantification); the disjunction '$\vee$' and the existential quantifier '$\exists$' are defined in the usual manner in terms of '$\neg$', '$\wedge$' and '$\forall$'.[1]

Given a language $\mathcal{L}$, we assume that each syntactical expression e from the vocabulary of $\mathcal{L}$ is (primitive recursively) assigned a fixed Gödel number or 'code' denoted by $\#e$. Let $\ulcorner e \urcorner$ denote the term (numeral) of $\mathcal{L}$ representing $\#e$; for a precise definition of the coding, we refer the readers to [17] for example. In formulating a system of truth, we also need to introduce the formal representations of some relations and operations which enable us to talk about the syntax of $\mathcal{L}$ within the system in question; it can be seen from our assumptions on the base system B and systems of truth that all the relations and functions (except the value function $\mathrm{val}(x)$) introduced below are primitive recursive, and we assume that they are represented by means of their corresponding primitive function symbols in $\mathcal{L}_0$.

For a given language $\mathcal{L}$, we first introduce some primitive recursive relations for its basic syntactical notions:

$\mathrm{Tm}_{\mathcal{L}}(x)$	$:\Leftrightarrow$	'x is a code of a term of $\mathcal{L}$';
$\mathrm{CT}_{\mathcal{L}}(x)$	$:\Leftrightarrow$	'x is a code of a closed term of $\mathcal{L}$';
$\mathrm{Var}_{\mathcal{L}}(x)$	$:\Leftrightarrow$	'x is a code of a variable of $\mathcal{L}$';
$\mathrm{AtFml}_{\mathcal{L}}(x)$	$:\Leftrightarrow$	'x is a code of an atomic formula of $\mathcal{L}$';
$\mathrm{AtSt}_{\mathcal{L}}(x)$	$:\Leftrightarrow$	'x is a code of an atomic sentence of $\mathcal{L}$';
$\mathrm{Fml}_{\mathcal{L}}(x)$	$:\Leftrightarrow$	'x is a code of a formula of $\mathcal{L}$';
$\mathrm{St}_{\mathcal{L}}(x)$	$:\Leftrightarrow$	'x is a code of a sentence of $\mathcal{L}$'.

[1] In the usual formulation of a classical first-order language, we only need to introduce the negation and conjunction (or disjunction). This is because other connectives can be defined in terms of these two and regarded as mere abbreviations; for example, '$A \to B$' is defined as '$\neg A \vee B$'. However, it is not necessarily the case in other logics such as intuitionistic logic, where '$\to$' and '$\wedge$' must be introduced as independent of '$\neg$' and '$\vee$'. In fact, in the case of the inner logic of Feferman's new theory DT [19], the conditional '$\to$' is not definable by '$\neg$' and '$\vee$'; namely, $T(\ulcorner \phi \to \psi \urcorner)$ and $T(\ulcorner \neg\phi \vee \psi \urcorner)$ are not equivalent in general. Thus, in order for us to treat DT and other systems altogether in the same framework, we introduce '$\to$' separately.

We will often suppress the subscript '$\mathcal{L}$' when it is clear from the context. In particular, since $\mathcal{L}$-terms and $\mathcal{L}$-variables are identical to those of $\mathcal{L}_{\mathsf{PA}}$, we always drop the subscript '$\mathcal{L}$' from '$\mathrm{CT}_{\mathcal{L}}$' and '$\mathrm{Var}_{\mathcal{L}}$'.

Since the base system extends PA, each language $\mathcal{L}$ of a system of truth has the (standard) primitive recursive representations of the following syntactical operations:

$$\mathrm{nm}(n) = \ulcorner n \urcorner \qquad \ulcorner t \urcorner \doteq \ulcorner s \urcorner = \ulcorner t = s \urcorner$$

$$\dot{\neg}\ulcorner \phi \urcorner = \ulcorner \neg \phi \urcorner \qquad \ulcorner \phi \urcorner \dot{\vee} \ulcorner \psi \urcorner = \ulcorner \phi \vee \psi \urcorner$$

$$\ulcorner \phi \urcorner \dot{\rightarrow} \ulcorner \psi \urcorner = \ulcorner \phi \rightarrow \psi \urcorner \qquad \dot{\forall}(\ulcorner v_k \urcorner, \ulcorner \phi \urcorner) = \ulcorner \forall v_k \phi \urcorner$$

$$\dot{R}(\ulcorner t_1 \urcorner, \ldots, \ulcorner t_k \urcorner) = \ulcorner R(t_0, \ldots, t_k) \urcorner \qquad \mathrm{sb}(\ulcorner e \urcorner, \ulcorner t \urcorner, \ulcorner v_k \urcorner) = \ulcorner e[t/v_k] \urcorner,$$

where $n \in \mathbb{N}$, t, s and $t_1, \ldots, t_k$ are terms, ϕ and ψ are formulae, v_k is the k-th variable, R is a predicate symbol of $\mathcal{L}$ with the arity k, and $e[t/v_k]$ is the expression obtained from an expression e by replacing each (free) occurrence of a variable v_k by a term t. We also need the 'value function' $\mathrm{val}(x)$ such that $\mathrm{val}(\ulcorner t \urcorner) = t$ for closed terms t: more precisely, $\mathrm{val}(x)$ is a function such that

$$\mathrm{val}(\ulcorner s \urcorner) := F\big(\mathrm{val}(s_0), \ldots, \mathrm{val}(s_k)\big)$$

provably in, say, PA, uniformly for each closed term s of the form $Fs_0 \ldots s_k$ where $s_0, \ldots, s_k$ are also closed terms and F is a $k+1$-ary function symbol. This value function $\mathrm{val}(x)$ can indeed be provably recursive in PA but not primitive recursive, since we could otherwise construct, for each $n \in \mathbb{N}$, a *primitive recursive* evaluation function $\mathrm{ev}_n(e, x_1, \ldots, x_n)$ of all the n-ary primitive recursive functions. We will occasionally write $\dot{x}$ and x° respectively for $\mathrm{nm}(x)$ and $\mathrm{val}(x)$ to save space.

For the sake of simplicity, we introduce some other abbreviations. First, we write $\dot{\forall} a.b$ for $\dot{\forall}(a, b)$

and $a(b/c)$ for $\mathrm{sb}\big(a, \mathrm{nm}(b), c\big)$; when it is clear from the context, we drop c and simply write $a(b)$. Second, for each formula $\phi(v)$, we write $\ulcorner \phi(\dot{x}) \urcorner$ for $\mathrm{sb}(\ulcorner \phi(v) \urcorner, \mathrm{nm}(x), \ulcorner v \urcorner)$, i.e., $\ulcorner \phi(v) \urcorner(x/\ulcorner v \urcorner)$. This definition is extended for multi-variable cases in an obvious manner. Third, we make use of the vector notation; e.g., we write $\ulcorner \phi(\dot{\vec{x}}) \urcorner$ for $\ulcorner \phi(\dot{x}_0, \ldots, \dot{x}_n) \urcorner$. Fourth, given an formula $\phi(\vec{x})$ where $\vec{x} = x_0, \ldots, x_n$ and $\vec{b} \in \mathrm{CT}$, we write $\mathrm{sb}_{\vec{x}}(\ulcorner \phi \urcorner, \vec{b})$ for

$$\mathrm{sb}\big(\mathrm{sb}\big(\cdots \mathrm{sb}\big(\mathrm{sb}(\ulcorner \phi \urcorner, b_0, \ulcorner v_0 \urcorner), b_1, \ulcorner v_1 \urcorner\big) \cdots\big), b_n, \ulcorner v_n \urcorner\big).$$

In the rest of the present chapter, we introduce and briefly discuss various systems of truth. We have fixed an arbitrary base system B, but when we need to mention a system with a specific base system C we will write $\mathsf{Q}[\![\mathsf{C}]\!]$ for a system of truth Q with a base system C: for example, $\mathsf{TB}[\![\mathsf{PA}]\!]$ is the system of Tarskian biconditionals (introduced later) with the base system PA.

2.2 Disquotational systems of truth I

Let us begin with the most naïve and oldest type of systems of truth: i.e., the so-called *disquotational* systems of truth. The background idea of disquotational systems of truth is that truth is to be axiomatized by Tarskian *T-biconditionals*, that is, equivalences of the form

$$\ulcorner \phi \urcorner \text{ is true iff } \phi,$$

which Tarski [82, 80] regarded as the 'material adequacy' condition for a definition of truth. As is well-known, if we allow ϕ to range over arbitrary sentences that may contain the truth predicate, the resulting system becomes inconsistent due to the Liar paradox. Hence, any disquotational approach to an axiomatization of truth must impose a certain restriction to the class of sentences over which ϕ above can range; we call such a class in a given system the *domain* of T-biconditionals.

The most straightforward and indeed oldest disquotational system is TB, which dates back to Tarski [82, pp. 255–257], though he himself rejected it because of its too weak deductive and expressive power as we shall discuss later.

Definition 2.2.1. TB is a system over $\mathcal{L}_T = \mathcal{L}_0 \cup \{T\}$ whose axioms consist of B-axioms and the TB-schema:

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

for each T-free sentence σ. Namely, the domain of T-biconditionals is restricted to T-free sentences in TB.

A natural expansion of T-biconditional is to allow parameters in ϕ: i.e., a schema of the form

$$\forall \vec{x}[\, T\ulcorner \phi(\dot{\vec{x}}) \urcorner \leftrightarrow \phi(\vec{x})\,].$$

This form of T-biconditional is called *uniform* T-biconditional.

Definition 2.2.2. UTB is a system over $\mathcal{L}_T$ whose axioms consist of B-axioms and the uniform T-biconditionals for all the T-free formulae.

The next proposition is well-known.

Proposition 2.2.3. TB and UTB are both proof-theoretically equivalent to B for $\mathcal{L}_0$.[2]

Here we use Feferman's notion of proof-theoretic reducibility and equivalence. For their precise definitions, we refer the readers to [16] or [18]; we will also discuss proof-theoretic reducibility in connection with relative truth definability in §3.2. We follow Feferman's notations for proof-theoretical reducibility and equivalence: i.e., $\mathsf{Q}_0 \leq \mathsf{Q}_1$ $[\Phi]$ means that Q_0 is proof-theoretically

[2]For the proof, see [37] or [48]. As a matter of fact, we can say more. Suppose B contains full induction for $\mathcal{L}_0$. Then, even if we add full induction for $\mathcal{L}_T$ to TB or UTB, the resulting system still remains conservative over B, since we can replace T by a definable partial truth predicate for a fixed complexity in each given derivation (see [33]). This idea is in fact what Tarski originally presented in [82]. However, since induction for partial truth predicates is not necessarily available in B in our current setting, this proof does not work for Proposition 2.2.3 in general.

reducible into Q_1 for a class of formulae Φ which is assumed to be a subset of $\mathcal{L}_{\mathsf{Q}_0} \cap \mathcal{L}_{\mathsf{Q}_1}$; $\mathsf{Q}_0 \equiv \mathsf{Q}_1$ $[\Phi]$ means that they are proof-theoretically equivalent for Φ, i.e., $\mathsf{Q}_0 \leq \mathsf{Q}_1$ $[\Phi]$ and $\mathsf{Q}_1 \leq \mathsf{Q}_0$ $[\Phi]$. In the following, when considering proof-theoretical reducibility for the base language $\mathcal{L}_0$, we will simply say 'proof-theoretically reducible' or 'proof-theoretically equivalent' without mentioning $\mathcal{L}_0$ and write $\mathsf{Q}_0 \leq (\equiv)\mathsf{Q}_1$ dropping the part '$[\mathcal{L}_0]$'.

From the standard proof of Proposition 2.2.3, we easily obtain:

Proposition 2.2.4. A system Q is said to be reflexive if Q proves the consistency of every finite subsystem of Q. Then, if B is reflexive, so are TB and UTB.

Tarski [82, p.257] observed that the principle of contradiction, i.e., $\forall x \in \mathrm{St}_{\mathcal{L}_0} \neg[T(x) \wedge T(\dot{\neg}x)]$, and similar general principles of truth (in the form of universal statements) are not derivable in TB; on the other hand, one can prove $\neg(T\ulcorner\sigma\urcorner \wedge T\ulcorner\neg\sigma\urcorner)$ for each single standard $\mathcal{L}_0$-sentence $\ulcorner\sigma\urcorner$. Of course, the principle of contradiction or other general principles can be added to TB as a new truth-theoretic axiom, but Tarski 'attach[es] little importance to this procedure' since 'it seems that every such enlargement of the axiom system has an accidental character' ([82, p.258]).

This attitude of Tarski is puzzling and seems to be incoherent with his Convention T that any definition of truth is formally correct and materially adequate whenever the T-schema is derivable from the definition. One immediate concern is that, by taking TB as the metasystem, we could define the truth of $\mathcal{L}_0$ in terms of the predicate T of $\mathcal{L}_T$ in TB. As to this concern, Tarski would probably respond that the predicate T is a semantical term and any definition by means of semantical terms is not taken into consideration for his theory; Tarski declared 'in this construction [of truth] I shall not make use of any semantical concept if I am not able to previously reduce it to other concepts ([82, p.153])' and thus Tarski was pursuing a reductive programme of defining away truth in terms of non-semantical terms. Tarski might have thought that any definition of truth in terms of non-semantical terms would yield a sufficient array of truth-theoretic principles. However, unfortunately, Mostowski [62] later provided a natural example of a truth definition in

purely non-semantical terms that is formally correct and materially adequate by Tarski's standards but does not imply general claims such as the principle of contradiction.[3]

Another objection Tarski raised against TB is that TB is not 'categorical'; in other words, TB does not 'unambiguously determine the extension of the symbol $[T]$ ([82, p.258])'. In view of this postulate 'categoricity', he requires systems of truth to meet the following: 'if we introduce into the [metasystem], alongside this symbol Tr, another primitive sign, e.g., the symbol 'Tr'' and set up analogous axioms for it, then the statement '$Tr = Tr'$' must be provable ([82, p.258]).' I will not discuss whether this postulate is necessary, but many of the subsequent systems satisfies this desideratum of Tarski indeed; in most cases, this is derivable due to full-induction.

2.3 Systems of Typed Truth

In his seminal paper [82], Tarski provided us with a means to define truth of a formal languages by means of the so-called 'inductive clauses' for truth. As is desired, his definition entails T-biconditionals for the target language. However, this definition cannot be carried out within the target language itself; this is his famous undefinability theorem of truth. These facts motivate the 'orthodox approach' [50] toward truth which leads to the Tarskian hierarchical view of languages and their truth. According to the 'orthodox approach', a language forms a certain hierarchy of its sublanguages in which the truth of one sublanguage is defined in and belongs to another sublanguage at a higher level in the hierarchy. Systems of hierarchical truth are motivated by this view and formulated so as to incorporate Tarski's 'inductive clauses' ('compositional axioms' in my terminology) for the truth of lower languages in their axioms.

Definition 2.3.1. The system TC^- over $\mathcal{L}_T$ consists of B-axioms and Tarski's 'inductive clauses' for truth:

[3] More precisely, Mostowski [62] defined a predicate in NBG which meets the T-schema for $\mathcal{L}_\in$.

T1. $\forall \vec{x} \in \mathrm{CT}[T(\dot{R}\vec{x}) \leftrightarrow R\vec{x}^{\circ}]$, for each $\mathcal{L}_0$-atomic R

T2. $\forall x \in \mathrm{St}_{\mathcal{L}_0}[T(\dot{\neg}x) \leftrightarrow \neg Tx]$

T3. $(\forall x \in \mathrm{St}_{\mathcal{L}_0})(\forall y \in \mathrm{St}_{\mathcal{L}_0})[T(x \dot{\wedge} y) \leftrightarrow (Tx \wedge Ty)]$

T4. $\forall x \in \mathrm{St}_{\mathcal{L}_0} \forall y \in \mathrm{St}_{\mathcal{L}_0}[T(x \dot{\rightarrow} y) \leftrightarrow (Tx \rightarrow Ty)]$

T5. $\forall x \forall y \big[\dot{\forall} x.y \in \mathrm{St}_{\mathcal{L}_0} \rightarrow \big(T(\dot{\forall} x.y) \leftrightarrow \forall z Ty(z/x)\big)\big]$.

The system TC is obtained by adding full induction for $\mathcal{L}_T$ to TC^-.

The acronym $\mathsf{TC}^{(-)}$ stands for Tarski's (inductive) clauses, and the above axioms are a direct axiomatization of Tarski's 'inductive clauses'. For example, the last axiom says that a sentence $\forall z \phi(z)$ is true iff $\phi(a)$ is true for each name a of an object in the domain. Note that the truth predicate characterized in this way is essentially the same thing as the so-called satisfaction class of models of arithmetic.[4]

Theorem 2.3.2 (Halbach [33])**.** TC^- is proof-theoretically equivalent to B.[5]

We say Q_0 is conservative over Q_1 for a set of formulae $\Phi \subset \mathcal{L}_{\mathsf{Q}_0} \cap \mathcal{L}_{\mathsf{Q}_1}$, written $\mathsf{Q}_0 \subset_\Phi \mathsf{Q}_1$, iff every Φ-theorem of Q_0 is derivable in Q_1 as well. Then it is immediately observed that conservativity for Φ follows from proof-theoretic reducibility for Φ. Hence, for example, it follows from Proposition 2.2.3 that $\mathsf{TB}[\![\mathsf{PA}]\!], \mathsf{UTB}[\![\mathsf{PA}]\!] \subset_{\mathcal{L}_0} \mathsf{PA}$: thus, they all have the same truth-free theorems. One more point is to be attended: if $\mathsf{Q}_0 \subset_{\mathcal{L}_0} \mathsf{Q}_1$ then the consistency of Q_1 implies that of Q_0, written $\mathsf{Q}_0 \leq_{\mathrm{Con}} \mathsf{Q}_1$.

On the other hand, TC is no longer conservative over B; indeed TC can prove the consistency of B (and thus TC^-). In the case of $\mathsf{B} = \mathsf{PA}$, it is known that:

[4] For a detailed exposition of satisfaction classes, see [47].

[5] Historically, Kotlarski, Krajewski and Lachlan [49] first gave a model-theoretic proof that $\mathsf{TC}^-[\![\mathsf{PA}]\!]$ is conservative over PA for Π^0_∞. Then, Halbach provided a finitary proof for this conservativity via cut-elimination and generalized this result to arbitrary base systems.

Theorem 2.3.3 (Feferman [23])**.** $\mathsf{TC}[\![\mathsf{PA}]\!]$ is proof-theoretically equivalent to ACA.

Proof. This match-up is established by the standard existence proof of a truth predicate of PA in ACA and the canonical embedding of second-order quantifications of ACA in $\mathsf{TC}[\![\mathsf{PA}]\!]$ using the inductive clauses; for more details see [37]. □

Note that since TC contains full induction for $\mathcal{L}_T$ and extends PA, it is reflexive (indeed, essentially reflexive) by Lemma III.3.47 of [29]. In addition, we can also see from Halbach's proof of Theorem 2.3.2 that we can uniformly find a primitive recursive function F such that any derivation of an $\mathcal{L}_0$-sentence from a finite set $\Delta \subset \mathsf{TC}^-$ can be primitive recursively transformed into a derivation of the same sentence from a finite set $F(\Delta) \subset \mathsf{B}$ (note that this says more than mere proof-theoretic reducibility). Hence, if B is reflexive then so is TC^- (but not essentially reflexive).[6]

Proposition 2.3.4. TC is reflexive, and if B is reflexive then so is TC^-.

TC^- and TC present how to axiomatize Tarski's 'inductive clauses'. Then, we can apply this to ascend up to higher and higher languages in the hierarchy of the 'orthodox approach'. Recall that the hierarchical view claims that no language can contain its own truth predicate and the truth predicate of a given language $\mathcal{L}$ is only contained and defined in a richer language $\mathcal{L}'$ than $\mathcal{L}$; then we iterate this process of ascending to the richer and richer languages and systems from the base system B and its language $\mathcal{L}_0$. The systems thus obtained form a hierarchy $\mathcal{L}_0, \mathcal{L}_1, \mathcal{L}_2, \cdots, \mathcal{L}_m, \cdots$. This hierarchy can go up beyond transfinite levels by taking the union of the lower languages at a limit stage. To formalize this idea, we first fix a standard notation system up to the least critical ordinal Γ_0[7]; we can of course consider higher levels, but Γ_0 is sufficient for our subsequent arguments. For the simplicity, we identify ordinals and their representations; e.g., $\bar{\alpha}$ denotes the numeral whose value is α in this notation system; but we often abuse notation and write α for $\bar{\alpha}$.

[6] As far as the I know, Proposition 2.3.4 was first explicitly noticed by Martin Fischer (by private communication).
[7] For its precise definition, see [11] or [66].

Definition 2.3.5. Let $\alpha \leq \Gamma_0$ and $\mathcal{L}_\alpha$ be a language $\mathcal{L}_0 \cup \{T_\beta\}_{\beta<\alpha}$ where T_β is a new unary predicate expressing the truth of β-th level. Then, the ramified truth system up to α, written as $\mathsf{RT}_{<\alpha}$, is a system over $\mathcal{L}_\alpha$ consisting of B-axioms, full-induction for $\mathcal{L}_\alpha$ and, for each $\gamma < \beta < \alpha$,

R1. $\forall \vec{x} \in \mathrm{CT}\big(T_\beta(\dot{R}\vec{x}) \leftrightarrow R\vec{x}^\circ\big)$, for each $\mathcal{L}_0$-atomic R.

R2. $\forall x \in \mathrm{St}_{\bar{\beta}}\big(T_\beta(\dot{\neg}x) \leftrightarrow \neg T_\beta(x)\big)$

R3. $\forall x, y \in \mathrm{St}_{\bar{\beta}}\big[\big(T_\beta(x \dot{\vee} y) \leftrightarrow T_\beta(x) \vee T_\beta(y)\big) \wedge \big(T_\beta(x \dot{\rightarrow} y) \leftrightarrow T_\beta(x) \rightarrow T_\beta(y)\big)\big]$

R4. $\forall z \forall x\big[\mathrm{St}_{\bar{\beta}}(\dot{\forall} z.x) \rightarrow \big(T_\beta(\dot{\forall} z.x) \leftrightarrow \forall y T_\beta x(y/z)\big)\big]$

R5. $\forall x \in \mathrm{CT}\big[\mathrm{St}_{\bar{\gamma}}(\mathrm{val}(x)) \rightarrow \big(T_\beta(\dot{T}_\gamma x) \leftrightarrow T_\gamma x^\circ\big)\big]$

R6. $\forall x \in \mathrm{CT} \forall \delta < \bar{\beta}\big[\mathrm{St}_\delta(\mathrm{val}(x)) \rightarrow \big(T_\beta(\dot{T}_\delta x) \leftrightarrow T_\beta x^\circ\big)\big]$,

where $\mathrm{St}_a(b)$ is a binary primitive recursive predicate on a and b such that $\mathrm{St}_a(b)$ iff $\mathrm{St}_{\mathcal{L}_\delta}(b)$ for $\delta = a$ and $\dot{T}_a b$ is a binary primitive recursive function on a and b whose value is $\ulcorner T_\beta t \urcorner$ when $a = \beta$ and $b = \ulcorner t \urcorner$.[8]

Then, the hierarchy of languages is represented by $\mathsf{RT}_{<\alpha}$ in the following way: a predicate T_β in $\mathcal{L}_{\beta+1}$ is the truth predicate of $\mathcal{L}_\beta$ in the sense that T_β satisfies the Tarski's inductive clauses **R1**-**R5** for $\mathcal{L}_\beta$; **R5** and **R6** express that the higher truth predicates include all the lower truth predicates and thus the higher languages are richer than the lower ones in the hierarchy; **R6** also indicates that T_λ for a limit λ is the union of all the lower truth predicates.

Suppose $\alpha < \beta$. Then, $\mathsf{RT}_{<\beta}$ can prove the consistency of $\mathsf{RT}_{<\alpha}$, since $\mathsf{RT}_{<\beta} \vdash \forall x \in \mathrm{St}_\alpha[\mathrm{Prv}_{\mathsf{RT}_{<\alpha}}(x) \rightarrow T_\alpha(x)]$. Hence, we have:

Proposition 2.3.6. For $\alpha < \beta$, $\mathsf{RT}_{<\alpha} \subsetneq_{\mathcal{L}_0} \mathsf{RT}_{<\beta}$.

[8] We follow the formulation of $\mathsf{RT}_{<\alpha}$ by Halbach in [37].

Let $\mathsf{RA}_{<\alpha}$ denote the ramified analysis up to α.[9] Then the ramified truth system and ramified analysis are related in the following way[10]:

Theorem 2.3.7 (Feferman [9])**.** $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\alpha}$ for each ordinal α.

For $\alpha < \Gamma_0$ and a formula $\phi(x)$ of a language $\mathcal{L}$, let $\mathrm{TI}(\alpha;\phi)$ denote

$$\forall x\big(\forall y < x\phi(y) \to \phi(x)\big) \to (\forall x \leq \alpha)\phi(x),$$

and $\mathrm{TI}_{\mathcal{L}}(\alpha)$ denote the schema $\mathrm{TI}(\alpha;\phi)$ for all $\phi \in \mathcal{L}$. Then, for $\alpha \leq \Gamma_0$, $\mathrm{TI}_{\mathcal{L}}(< \alpha)$ denotes transfinite induction up to α: i.e., $\mathrm{TI}_{\mathcal{L}}(\beta)$ for all $\beta < \alpha$. It is known that, for an epsilon number α (i.e., $\alpha = \omega^\alpha$), $\mathsf{RA}_{<\alpha}$ proves transfinite induction up to $\varphi_\alpha 0$ for arithmetical formulae, i.e., $\mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(< \varphi_\alpha 0)$, where φ is the binary Veblen function. It follows from the equivalence between $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!]$ and $\mathsf{RA}_{<\alpha}$ that $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!]$ proves the same transfinite induction (cf. [11, 17]). Then, the proof of $\mathsf{RT}_{<\alpha}[\![\mathsf{PA}]\!] \vdash \mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(< \varphi_\alpha 0)$ induced from the proof of $\mathsf{RA}_{<\alpha} \vdash \mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(< \varphi_\alpha 0)$ can be straightforwardly transformed to the proof of $\mathsf{RT}_{<\alpha}[\![\mathsf{B}]\!] \vdash \mathrm{TI}_{\mathcal{L}_{\mathsf{B}}}(< \varphi_\alpha 0)$ for an arbitrary base system B which meets the conditions we have posed. Thus, we have the next theorem.

Theorem 2.3.8. Let α be an epsilon number. Then, $\mathsf{RT}_{<\alpha} \vdash \mathrm{TI}_{\mathcal{L}_0}(< \varphi_\alpha 0)$.

2.4 Systems of iterative compositional truth I

The iterative compositional conception of truth is best represented by the following passage of Feferman [17, p.18]: 'truth or falsity is *grounded* in atomic facts from the base language $\mathcal{L}$, i.e., can be determined from such facts by evaluation according to the rules of truth for the connectives and quantifiers, and where statements of the form $[T(\ulcorner A \urcorner)]$ are evaluated to be true (false) only when A itself has already been verified (falsified).' It is 'iterative' because a sentence $T(\ulcorner A \urcorner)$ is true only

[9] For the precise definition, see [10] or [17].
[10] Feferman gave a sketch of its proof in [9]. A more detailed proof can be found in [32].

if A is already true, and it is 'compositional' because a compositional sentence is determined only by its components according to the evaluation rule for connectives and quantifiers. Kripke [50] first focused on this iterative compositional conception of truth and gave a model-theoretic formulation of such truth, and then Feferman [17] later gave an axiomatic formalization KF of this conception.[11] For simplicity, we write $F(x)$ for $T(\dot{\neg}x)$ (its intended meaning is 'false') in the following.

Definition 2.4.1. The system KF^- over $\mathcal{L}_T$ consists of B and:

K0 $\forall z \in \mathrm{Var} \forall y \in \mathrm{CT} \forall x \big[\mathrm{St}_{\mathcal{L}_T}(\dot{\forall} z.x) \to \big(T\mathrm{sb}(x,y,z) \leftrightarrow T\mathrm{sb}\big(x, \mathrm{nm}(y^\circ), z\big)\big)\big]$

K1 For each atomic R of $\mathcal{L}_0$,

$$\forall \vec{x} \in \mathrm{CT}\big(T(\dot{R}\vec{x}) \leftrightarrow R\vec{x}^\circ\big)\big] \wedge \forall \vec{x} \in \mathrm{CT}\big(F(\dot{R}\vec{x}) \leftrightarrow \neg R\vec{x}^\circ\big)\big]$$

K2 $\forall x \in \mathrm{CT}[T(\dot{T}x) \leftrightarrow Tx^\circ] \wedge \forall x \in \mathrm{CT}[F(\dot{T}x) \leftrightarrow Fx^\circ]$

K3 $\forall x \in \mathrm{St}_{\mathcal{L}_T}\big[T(\dot{\neg}\dot{\neg}x) \leftrightarrow Tx\big]$

K4 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T}\big[T(x \dot{\wedge} y) \leftrightarrow (Tx \wedge Ty)\big]$

K5 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T}\big[F(x \dot{\wedge} y) \leftrightarrow (Fx \vee Fy)\big]$

K6 $\forall z \forall x \big[\mathrm{St}_{\mathcal{L}_T}(\dot{\forall} z.x) \to \big(T(\dot{\forall} z.x) \leftrightarrow \forall y Tx(y/z)\big)\big]$

K7 $\forall z \forall x \big[\mathrm{St}_{\mathcal{L}_T}(\dot{\forall} z.x) \to \big(F(\dot{\forall} z.x) \leftrightarrow \exists y Fx(y/z)\big)\big]$

K8 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T} \to \big[\big(T(x \dot{\to} y) \leftrightarrow T(\dot{\neg}x \dot{\vee} y)\big) \wedge \big(F(x \dot{\to} y) \leftrightarrow F(\dot{\neg}x \dot{\vee} y)\big)\big]$[12]

Then, KF is defined as KF^- plus full induction for $\mathcal{L}_T$.

The axioms of KF is regarded as a partial axiomatization of (closed off) Kripkean fixed-point models with Strong Kleene schema. As is known, Strong Kleene Logic is the three valued logic with the following truth table:

[11] The system KF and the main results about it were in fact presented by Feferman in 1979. For more historical issues, see [30].

[12] This axiom is not usually postulated in the literature, but we need it since we have assumed that '$\to$' is a primitive logical connective and since '$\to$' is definable by '$\neg$' and '$\vee$' in Strong Kleene Logic.

$\neg$	
T	F
U	U
F	T

$\wedge$	T	U	F
T	T	U	F
U	U	U	F
F	F	F	F

The conditional '$\rightarrow$' and conjunction '$\wedge$' are defined in the usual manner in terms of '$\neg$' and '$\vee$'. The existential and universal quantifiers are defined as infinitary disjunction and conjunction respectively.

Now it is clear how KF represents the iterative compositional conception of truth with the Strong Kleene schema. For example, **K5** represents how falsity is assigned to a disjunctive sentence according to the Strong Kleene schema: that is, a disjunctive sentence $\phi \vee \psi$ is false iff both conjuncts are false. For another example, **K2** represents the iterative character of truth in which a sentence $T(\phi)$ is true (false) iff ϕ is true (false). The axiom **K0** is postulated to assure that the truth of a given sentence $\phi(t)$ depends only on the value of t.[13]

The proof-theoretic strength of KF and KF^- are known:

Theorem 2.4.2 (Feferman [17])**.** $\mathsf{KF}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\varepsilon_0}$. Thus $\mathsf{KF}[\![\mathsf{PA}]\!]$ is also equivalent to, e.g., Σ^1_1-AC and $\widehat{\mathsf{ID}}_1$.[14]

It is known that $\mathsf{KF}^-[\![\mathsf{PA}]\!] \equiv \mathsf{PA}$. To my knowledge, this equivalence was first observed by Cantini [5, Theorem 9.10]. Cantini showed that $\mathsf{KF}_{\mathrm{tot}}$, an extension of $\mathsf{KF}^-[\![\mathsf{PA}]\!]$, is reducible into PA. His proof does apply for any base system B which meets the condition we have already put in the last subsection. We show an alternative proof of the equivalence $\mathsf{KF}^-[\![\mathsf{PA}]\!] \equiv \mathsf{PA}$. The proof is a straightforward modification of Halbach's conservativity proof of $\mathsf{TC}^-[\![\mathsf{B}]\!]$ over $\mathsf{PA}[\![\mathsf{B}]\!]$.

Theorem 2.4.3. $\mathsf{KF}^- \equiv \mathsf{B}$.

[13] **K0** was dropped in Feferman's original formulation of KF, since it is derivable from the other axioms in KF. However, it is not derivable in KF^- without full induction for $\mathcal{L}_T$ and thus introduced by Cantini for KF^-. It is sometimes called the regularity axiom.

[14] For the precise definitions of Σ^1_1-AC and $\widehat{\mathsf{ID}}_1$, see [14].

Proof. We first reformulate KF^- in a sequent calculus à la Tait. The axioms for the truth predicates (i.e., **K1**-**K8**) are reformulated as the following inference rules. For a technical reason, we drop **K0** and the following in fact shows that $\mathsf{KF}^- - \mathbf{K0} \leq \mathsf{B}$.[15]

$$\frac{\Gamma, R(\vec{s}^\circ)}{\Gamma, \neg\mathrm{CT}(\vec{s}), s \neq \dot{R}\vec{s}, Tt}\ \mathbf{K1}^+ \qquad \frac{\Gamma, \neg R(\vec{s}^\circ)}{\Gamma, \neg\mathrm{CT}(\vec{s}), s \neq \dot{R}\vec{s}, \neg Tt}\ \mathbf{K1}^-$$

$$\frac{\Gamma, Ts^\circ}{\Gamma, \neg\mathrm{CT}(s), t \neq \dot{T}s, Tt}\ \mathbf{K2\text{-}1}^+ \qquad \frac{\Gamma, \neg Ts^\circ}{\Gamma, \neg\mathrm{CT}(s), t \neq \dot{T}s, \neg T(t)}\ \mathbf{K2\text{-}1}^-$$

$$\frac{\Gamma, Fs^\circ}{\Gamma, \neg\mathrm{CT}(s), t \neq \dot{T}s, Ft}\ \mathbf{K2\text{-}2}^+ \qquad \frac{\Gamma, \neg Fs^\circ}{\Gamma, \neg\mathrm{CT}(s), t \neq \dot{T}s, \neg F(t)}\ \mathbf{K2\text{-}2}^-$$

$$\frac{\Gamma, Ts}{\Gamma, \neg\mathrm{St}_\beta(t), t \neq \dot{\neg}\dot{\neg}s, Tt}\ \mathbf{K3}^+ \qquad \frac{\Gamma, \neg Ts}{\Gamma, \neg\mathrm{St}_\beta(t), t \neq \dot{\neg}\dot{\neg}s, \neg Tt}\ \mathbf{K3}^-$$

$$\frac{\Gamma, Ts \wedge Tt}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq s\dot{\wedge}t, Tr}\ \mathbf{K4}^+ \qquad \frac{\Gamma, \neg Ts \vee \neg Tt}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq s\dot{\wedge}t, \neg Tr}\ \mathbf{K4}^-$$

$$\frac{\Gamma, Fs \vee Ft}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq s\dot{\wedge}t, Fr}\ \mathbf{K5}^+ \qquad \frac{\Gamma, \neg Fs \wedge \neg Ft}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq s\dot{\wedge}t, \neg Fr}\ \mathbf{K5}^-$$

$$\frac{\Gamma, \forall y Ts(y)}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq \dot{\forall}t.s, Tr}\ \mathbf{K6}^+ \qquad \frac{\Gamma, \neg\forall y Ts(y)}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq \dot{\forall}t.s, \neg Tr}\ \mathbf{K6}^-$$

$$\frac{\Gamma, \exists y Fs(y)}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq \dot{\forall}t.s, Fr}\ \mathbf{K7}^+ \qquad \frac{\Gamma, \neg\exists y Ts(y)}{\Gamma, \neg\mathrm{St}_\beta(r), r \neq \dot{\forall}t.s, \neg Fr}\ \mathbf{K7}^-$$

for any terms s and t. Let us call these T-rules. The logical rules and identity rules are as usual. Each derivation is represented by a finite tree.

We will show that, given a derivation $\mathcal{D}$, we can effectively construct a derivation $\mathcal{D}'$ with the same conclusion which contains applications of cut only to truth-free formulae. The main induction is on the number n of cut to formulae with truth predicates. Suppose we have shown the claim until $n = m+1$. Then we can focus on a subderivation $\mathcal{D}_0$ of $\mathcal{D}$ which contains only one application of cut to formulae with truth predicates and ends with it. If we can eliminate this cut from $\mathcal{D}_0$,

[15] Dropping **K0** causes no loss of generality for the current purpose since we have $\mathsf{KF}^- \leq \mathsf{KF}^- - \mathbf{K0}$. This can be shown by interpreting $T(\ulcorner\phi(\vec{x})\urcorner)$ by $T\big(\mathrm{sb}(\ulcorner\phi\urcorner, \mathrm{nm} \circ \mathrm{val}(\vec{x}), \vec{x}\big)$ in $\mathsf{KF}^- - \mathbf{K0}$.

our claim for n follows from IH.

Then, we define the T-*complexity* of this derivation $\mathcal{D}_0$. The T-complexity of a formula occurring in a sequent of $\mathcal{D}_0$ is the numbers of T-rules used to obtain it: more precisely, the T-complexity of a truth-free formula is always 0; the first occurrence of a formula ϕ (thus introduced by identity or logical axioms) containing truth predicates is assigned T-complexity 1; the T-complexities of side formulae of inferences are not changed at all; if ϕ is obtained by a T-rule from ψ then the T-complexity of ϕ is that of ψ plus 1; if ϕ is obtained by a logical rules from ψ_i ($i \in I$) then the T-complexity of ϕ is the maximum of the T-complexities of ψ_i's. For more detailed explanation, see [33]. Then, the T-complexity of $\mathcal{D}_0$ is defined to be the sum of T-complexity of the critical formula ϕ and $\neg\phi$ of the last inference (i.e., cut) of $\mathcal{D}_0$. We will show that $\mathcal{D}_0$ can be transformed into a derivation $\mathcal{D}_1$ with the same conclusion but with less T-complexity; our claim is obtained thus by induction on T-complexity of $\mathcal{D}_0$.

Given $\mathcal{D}_0$, in the usual manner, we can transform $\mathcal{D}_0$ without changing its T-complexity into an derivation $\mathcal{D}_0'$ in which cut is only applied to atomic formulae (i.e., the cut-rank is 1) and in which the critical formulae of cut are introduced just in the premise of the critical sequents (i.e., the sum of the left-rank and the right-rank is 2 in the terminology of [79]). Hence, we can assume without loss of generality that the last inference of $\mathcal{D}_0$ is of the form:

$$\dfrac{\dfrac{\dfrac{\vdots}{\Gamma', \Phi}}{\Gamma, Tt} \qquad \dfrac{\dfrac{\vdots}{\Gamma'', \Psi}}{\Gamma, \neg Tt}}{\Gamma} \tag{2.1}$$

For the base step, suppose Γ, Tt is introduced by an axiom. We only consider the essential case in which it is obtained by the identity axiom. This can be shown by cases according to how $\Gamma, \neg Tt$ is obtained, and we show only one case in which it is obtained by $\mathbf{K2^{-}2^{-}}$. Then, the last inference

is of the form:

$$\dfrac{\Gamma_0, s \neq t, \neg Ts, Tt \qquad \dfrac{\dfrac{\vdots}{\Gamma_1, \neg Tr^\circ}}{\Gamma_1, \neg \mathrm{CT}(r), t \neq \dot{T}r, \neg Tt}}{\Delta}$$

Let Γ_1' be $\Gamma_1 \cup \{t \neq \dot{T}r\}$. Then, for a suitable Γ_0', we can derive

$$\dfrac{\Gamma_0', s \neq t, \dot{T}r \neq t, \dot{T}r = s \qquad \dfrac{\Gamma_1', \neg Tr^\circ}{\Gamma_1', \neg \mathrm{CT}(r), \dot{T}r \neq s, \neg Ts}}{\Delta}$$

by applying cut to $\dot{T}_\beta s = r$ and $\dot{T}_\beta s \neq r$. This derivation contains no cut to formulae with truth predicates.

For the induction step, suppose Γ, Tt and $\Gamma, \neg Tt$ are obtained non-dual T-rules. Then, due to the unique readability, Γ can be obtained without cut to any formula containing truth predicates. For example, if they are obtained by $\mathbf{K3}^+$ and $\mathbf{K4}^-$, then Γ contains $t \neq \dot{\neg}\dot{\neg}t_0$ and $t \neq t_1 \dot{\wedge} t_2$ for some terms t_0, t_1, t_2 but the sequence $t \neq \dot{\neg}\dot{\neg}t_0$, $t \neq t_1 \dot{\wedge} t_2$ is derivable in B by the unique readability. Assume Γ is obtained by some dual T-rules. Then, it must be of the form

$$\dfrac{\dfrac{\dfrac{\vdots}{\Gamma', \Phi}}{\Gamma, Tt} \qquad \dfrac{\dfrac{\vdots}{\Gamma', \neg\Phi}}{\Gamma, \neg Tt}}{\Gamma}$$

which is transformed into a derivation with less T-complexity by applying cut to Φ and $\neg\Phi$. We have completed the proof. □

As we have mentioned, Cantini showed that a supersystem $\mathsf{KF}_{\mathrm{tot}}$ of KF^- is proof-theoretically reducible into PA. However, $\mathsf{KF}_{\mathrm{tot}}$ is not reflexive and indeed his proof does not directly yield the

reflexivity of $\mathsf{KF}^-[\![\mathsf{PA}]\!]$, since the bound of complexity of the reduced proof in PA depends on the length of the original proof in $\mathsf{KF}_{\mathrm{tot}}$. In contrast, it is observed that the above proof of $\mathsf{KF}^- \equiv \mathsf{B}$ yields the reflexivity result in the same manner as Proposition 2.3.4.

Proposition 2.4.4. KF is (essentially) reflexive. If B is reflexive, then so is KF^-.

In the context of systems of iterative compositional truth, we sometimes consider the two special axioms: the consistency axiom and completeness axiom which are defined respectively by

$$\mathrm{Cons} : \forall x \in \mathrm{St}_{\mathcal{L}_T} \neg(Tx \wedge Fx) \qquad\qquad \mathrm{Comp} : \forall x \in \mathrm{St}_{\mathcal{L}_T}(Tx \vee Fx).$$

Feferman [17] argued that the character of Cons is different from the rules of evaluation incorporated in KF and his argument should apply to Comp as well.[16] In fact, neither of them is derivable from KF and adding both to KF leads to a contradiction.[17] On the other hand, Cantini [5] showed that either (but not both) of them can be consistently and even conservatively (for $\mathcal{L}_0$) added to KF. We will see in §2.5 and §5 that Cons is particularly important in the connection between iterative compositional systems and determinate systems of truth.

Concluding this subsection, we have a closer look at iterativity and compositionality. We can now extract from the above observations more general characteristics of the iterative compositional systems of truth. That is, they may be characterized in the following way: for each logical connectives C, its axioms for compositional truth are of the form

$$\forall \vec{x}\big[\mathrm{St}_{\mathcal{L}_T}(x_0) \wedge \ldots \wedge \ldots, \mathrm{St}_{\mathcal{L}_T}(x_k) \rightarrow \big(T(\dot{C}\vec{x}) \leftrightarrow R_C(\vec{x})\big)\big] \tag{2.2}$$

for a certain formula R_C suitable for a given evaluation rule for C; the axioms for quantifiers are of

[16] Feferman denotes Cons by *Disj* there. The notations Cons and Comp are due to Cantini [5], but, to my knowledge, these two axioms first appears in [23].

[17] Given a least Kripkean fixed-point model M with Strong Kleene schema, if M is closed off in the way that all the undetermined sentences are pushed into the *extension* of T, then the resulting classical structure is a model of KF + Comp; if M is closed-off in the usual manner then it is a model of KF + Cons.

the parallel forms for each quantifier Q and its evaluation formula R_Q; the axioms for iterativity are of the form like $T(TA) \leftrightarrow TA$ as in KF. Although it is highly debatable what kinds of evaluation formulae R_C and R_Q can be regarded as 'compositional', any system for iterative compositional truth should be axiomatized in this way. Now, let S be an iterative compositional system of truth with the full induction, and suppose that R_C and R_Q are T-positive for each connective C and quantifier Q. Then, we can show that the truth of S can be constructed as a fixed-point of a positive arithmetical operator. Therefore, when the base system B is PA, S is proof-theoretically reducible to Σ^1_1-AC via Aczel's construction of a Σ^1_1 fixed-point of a Σ^1_1-formula (see [17]).

2.5 Disquotational systems of Truth II

The domains of T-biconditionals are restricted to arithmetical sentences in TB and UTB, but one can consider a disquotational system of truth whose domain of T-biconditionals is larger and may contain sentences with T. However, it has turned out to be difficult to specify such a domain in a natural and consistent way among uncountably many possibilities.[18] As far as the I know, only one such expansion of the domain has so far been presented; the schema PUTB is the uniform T-biconditionals with the domain the set of sentences in which T occurs only positively. The acronym PUTB stands for *positive uniform T-biconditionals*.[19]

The schema PUTB properly extends UTB, but if we add PUTB to B, it is known that the resulting system is still conservative over B. On the other hand, if we add the full induction for $\mathcal{L}_T$ together with PUTB, the resulting system is no longer conservative; we call this system PUTB.

[18] McGee [59] showed that there are uncountably many different and incompatible maximal consistent sets of T-biconditionals; to make the matter worse, Cieśliński [8] further showed that there are uncountably many different and incompatible maximal sets of T-biconditionals that are even *conservative over the base system*. It is also observed by Halbach (by private communication) that any system can be equivalently reformulated in a purely disquotational form by means of McGee's trick [59]. Thus the strength of purely disquotational systems can range over all the possible ones.

[19] As far as I know, the PUTB schema was first considered by Cantini [5], and Halbach [36] used it as a formal system of truth. The appellation PUTB is due to Halbach.

Definition 2.5.1. Given a $\mathcal{L}_T$-formula ϕ, ϕ is said to be T-positive if every occurrence of T in ϕ is positive. The system PUTB over $\mathcal{L}_T$ consists of B-axioms, full induction for $\mathcal{L}_T$ and the schema PUTB: for each T-positive formula $\phi(\vec{x})$,

$$\forall \vec{x}[T(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \leftrightarrow \phi(\vec{x})].$$

Then, PUTB^- denotes the system B plus the PUTB-schema only (without full induction for $\mathcal{L}_T$); PUTB^- is proof-theoretically equivalent to B as was noted.

We shall see later that PUTB is fairly strong system.

It is also known that PUTB is a subsystem of KF; more precisely, we have:

Lemma 2.5.2 (Cantini [5])**.** KF^- proves the PUTB schema. Hence PUTB^- (or PUTB) is a subsystem of KF^- (KF, respectively).

2.6 Systems of iterative compositional truth 2

KF was an axiomatic formulation of Kripke's fixed-point definition of self-applicable truth with the Strong Kleene schema. It is then natural to consider 'KF-style' axiomatic formulation of Kripke's fixed-point truth definition *with other schemata*. As was suggested by Feferman [17], one natural and straightforward alternative is the Weak Kleene schema, which is also a three-valued evaluation schema named after Kleene. Its evaluation for the logical connectives is given as follows:

$\neg$	
T	F
U	U
F	T

$\wedge$	T	U	F
T	T	U	F
U	U	U	U
F	F	U	F

The universal quantifier is interpreted as infinitary (Weak Kleene) conjunction: i.e., $\forall x \phi(x)$ is true iff $\phi(x)$ is true for all x; $\forall x \phi$ is false iff $\phi(x)$ is false for some x and $\phi(x)$ is either true or false for

all x. The other logical primitives are defined in the usual manner from them.

Then, the system WKF, 'KF-like' system of self applicable truth with the Weak Kleene schema, is defined as follows.

Definition 2.6.1. WKF^{-} consists of all the axioms of PA, **K0**–**K4**, **K6**, **K8** and the following axioms (instead of **K5** and **K7**):

WK5 $\mathrm{St}_{\mathcal{L}_T}(x) \wedge \mathrm{St}_{\mathcal{L}_T} \rightarrow \Big[F(x \dot{\wedge} y) \leftrightarrow \Big((Fx \wedge Fy) \vee (Fx \wedge Ty) \vee (Tx \wedge Fy)\Big)\Big]$

WK7 $\mathrm{Var}(z) \wedge \mathrm{St}_{\mathcal{L}_T}(\dot{\forall} z.x) \rightarrow \Big[F(\dot{\forall} z.x) \leftrightarrow \Big(\forall y\big[Tx(y/z) \vee Fx(y/z))\big] \wedge \exists y Fx(y/z)\Big)\Big]$

Then WKF plus full induction for $\mathcal{L}_T$.

These axioms represent how compound sentences are evaluated according to the Weak Kleene schema in the form of (2.2) in §2.4; for example, **WK5** says that a compound sentence $A \wedge B$ is evaluated to be false when both A and B has a definite truth value and either of them is false.

Feferman [19] introduced a new type of system DT of self-applicable truth. In my [20], I pointed out that his DT is in fact a 'disguised' KF-like system with a special evaluation schema. I call this schema *Feferman schema* and the logic based on this schema *Feferman Logic*. I will introduce DT itself in the next section, but I consider that the KF-like system with Feferman schema below in the present section; the connection between it and DT will be explained later in the next section.

Feferman schema evaluates the connectives and quantifiers in the same way as the Weak Kleene one only except for the conditional $\rightarrow$. The evaluation rule for the conditional '$\rightarrow$' is given by the following:

$\rightarrow$	T	U	F
T	T	U	F
U	U	U	U
F	T	T	T

Then, the 'KF-like' system FKF with Feferman schema is defined as follows.

Definition 2.6.2. FKF$^-$ consists of all the axioms of PA, WKF$^-$ minus **K8**, and the following:

FK9 $\mathrm{St}_{\mathcal{L}}(x) \wedge \mathrm{St}_{\mathcal{L}}(y) \to \big[T(x \dot{\to} y) \leftrightarrow \big((Tx \wedge Ty) \vee Fx\big)\big]$

FK10 $\mathrm{St}_{\mathcal{L}}(x) \wedge \mathrm{St}_{\mathcal{L}}(y) \to \big[F(x \dot{\to} y) \leftrightarrow (Tx \wedge Fy)\big]$.

We will briefly review some basic results of $\mathsf{KF}^{(-)}$, $\mathsf{FKF}^{(-)}$, $\mathsf{WKF}^{(-)}$ and $\mathsf{PUTB}^{(-)}$ below.

Proposition 2.6.3. KF$^-$, FKF^-, WKF^- and PUTB^- prove the following schemata:

(i) $\forall \vec{x}\big[T\big(\ulcorner \phi(\dot{\vec{x}}) \urcorner\big) \leftrightarrow \phi(\vec{x})\big]$, if ϕ contains no occurrence of T.

(ii) KF$^- \vdash D^+(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \to \big[T\ulcorner \phi(\dot{\vec{x}}) \urcorner \leftrightarrow \phi(\vec{x})\big]$, where $D^+(x)$ denotes $(Tx \vee Fx) \wedge \neg(Tx \wedge Fx)$; note that $D^+(x)$ is equivalent to $F(x) \leftrightarrow \neg T(x)$.

(iii) $\mathrm{Cons} \to \forall \vec{x}\big[T\big(\ulcorner \phi(\dot{\vec{x}}) \urcorner\big) \to \phi(\vec{x})\big]$, for an arbitrary formula ϕ.

(iv) $\mathrm{Comp} \to \forall \vec{x}\big[\phi(\vec{x}) \to T\big(\ulcorner \phi(\dot{\vec{x}}) \urcorner\big)\big]$, for an arbitrary formula ϕ.

Remark 5. As was seen in Lemma 2.5.2, the schema PUTB is derivable in KF^-. However, neither FKF nor WKF derives it. Let λ denote the Liar sentence: the sentence such that $\ulcorner F\ulcorner \lambda \urcorner \urcorner = \ulcorner \lambda \urcorner$ provably, say, in PA. Then, $\mathrm{FKF}\big(\mathrm{WKF}\big) \vdash 0 = 0 \vee \lambda$, but $\mathrm{FKF}\big(\mathrm{WKF}\big) \nvdash T(\ulcorner 0 = 0 \vee \lambda \urcorner)$, as will be shown later in Chapter 4.

Proposition 2.6.4. Let λ be the above liar sentence and Q be either WKF^-, FKF^-, KF^- or PUTB^-. Then, $\mathsf{Q} + \mathrm{Cons} \vdash \neg\lambda$ and $\mathsf{Q} + \mathrm{Comp} \vdash \lambda$.

Define a translation ^{c} as follows: $\phi^c = \phi$ when ϕ is $\mathcal{L}_0$-atomic, $\phi^c = \neg F(t)$ when ϕ is Tt, and ^{c} commutes with the connectives and quantifiers.

Proposition 2.6.5 (Cantini [5])**.** If $\mathsf{KF}^{(-)} \vdash \phi$ then $\mathsf{KF}^{(-)} \vdash \phi^c$. As a consequence, $\mathsf{KF}^{(-)} + \mathrm{Cons} \vdash \phi$ implies $\mathsf{KF}^{(-)} + \mathrm{Comp} \vdash \phi^c$ and *vice versa*.

Proof. This is proved by showing that $\mathsf{KF}^- \vdash \sigma^c$ for each KF^--axiom σ. □

By contrast, as we shall see in Chapter 4, this duality fails for $\mathsf{WKF}^{(-)}$, $\mathsf{FKF}^{(-)}$ and $\mathsf{PUTB}^{(-)}$.

2.7 Systems of Determinate Truth

Feferman's system DT [19] brought about a new form of axiomatic systems of truth. The background idea of DT, which Feferman ascribes to Russell, is that every predicate R has a domain D of significance and it makes sense to apply R only to objects in D and therefore the principles which are supposed to characterize the concept expressed by R are to be applied only to objects in D. Thus the system of truth based on this view would have two predicates T and D representing truth and its domain of significance. According to Feferman, such a domain of significance in the case of truth consists of *meaningful and determinate* sentences, i.e., sentences with a definite truth value, true or false, and the T-schema and other principles characterizing truth can be restricted to such sentences. Feferman's argument for this identification of the domain with the set of determinate sentences is simple. Let D be the domain of truth predicate T. Then, we have $[T(A) \vee F(A)] \rightarrow D(A)$.[20] Conversely, since the T-biconditional, $T(A) \leftrightarrow A$ and $F(A) \leftrightarrow \neg A$, obtains for every A in D, we have $D(A) \rightarrow [T(A) \vee F(A) \leftrightarrow (A \vee \neg A)]$; thus, we have $D(A) \rightarrow [T(A) \vee F(A)]$. This argument is intuitively convincing and Feferman just identifies $D(x)$ with $T(x) \vee F(x)$; thus the new predicate D becomes redundant.[21] In addition, Feferman further requires that D must be *strongly compositional* in the sense that a compound sentence is in D iff all the substitution instances of its subformulae by meaningful terms belong to D.[22] Feferman presented a system DT to capture these views on truth.

Definition 2.7.1. DT^{-} is a system over $\mathcal{L}_T$ whose axioms consist of B-axioms, **K0** and:

D1 $\forall x[\mathrm{AtSent}_{\mathcal{L}_0}(x) \rightarrow D(x)]$

[20] To be precise, in this argument, Feferman implicitly assumes that $D(\neg A) \rightarrow D(A)$.

[21] Formally speaking, there is a subtle matter about this argument. The argument to conclude $D(A) \leftrightarrow [T(A) \vee F(A)]$ for each (standard) sentence A does not formally entail $D(x) \leftrightarrow [T(x) \vee F(x)]$ for all x. This is because any (recursive) arithmetical system ($\supset \mathsf{PA}$) must have a nonstandard model M and thus nonstandard sentences in M by the usual overspill argument. Hence, when one introduces a predicate D separately from T, the equivalence $D(x) \leftrightarrow T(x) \vee F(x)$ is not necessarily *provable*. It may be of some interest to consider and investigate a DT-like system in which D and T are separately introduced.

[22] As Feferman himself pointed out, strong compositionality in this sense is not met with respect to the conditional '$\rightarrow$' due to **D11**. If we pose full strong compositionality, the resulting system becomes identical to the system $\mathsf{WKF}+\mathrm{Cons}$ which will be introduced later. See §5.

D2 $\forall x \in \mathrm{CT}\big[D\big(\dot{T}x\big) \leftrightarrow D\big(\mathrm{val}(x)\big)\big]$

D3 $\forall x \in \mathrm{St}_{\mathcal{L}_T}\big[D(\dot{\neg}x) \leftrightarrow D(x)\big]$

D4 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T}\big[D(x\dot{\wedge}y) \leftrightarrow (Dx \wedge Dy)\big]$

D5 $\forall z \forall x\big[\mathrm{St}_{\mathcal{L}_T}(\dot{\forall}z.x) \rightarrow \big(D(\dot{\forall}z.x) \leftrightarrow \forall y Dx(y/z)\big)\big]$

D6 $\forall x \in \mathrm{CT}\big[T(\dot{R}\vec{x}) \leftrightarrow R\big(\mathrm{val}(\vec{x})\big)\big]$, for each atomic R of $\mathcal{L}_0$

D7 $\forall x \in \mathrm{CT}\big[D\big(\mathrm{val}(x)\big) \rightarrow \big(T(\dot{T}x) \leftrightarrow T(\mathrm{val}(x))\big)\big]$

D8 $\forall x \in \mathrm{St}_{\mathcal{L}_T}\big[D(x) \rightarrow \big(T(\dot{\neg}x) \leftrightarrow \neg Tx\big)\big]$

D9 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T}\big[D(x\dot{\vee}y) \rightarrow \big(T(x\dot{\wedge}y) \leftrightarrow (Tx \wedge Ty)\big)\big]$

D10 $\forall z \forall x\big[\mathrm{St}_{\mathcal{L}_T}(\dot{\forall}z.x) \wedge D(\dot{\forall}z.x) \rightarrow \big(T(\dot{\forall}z.x) \leftrightarrow \forall y Tx(y/z)\big)\big]$

D11 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T}\big[D(x\dot{\rightarrow}y) \leftrightarrow \big(Dx \wedge (Tx \rightarrow Dy)\big)\big]$

D12 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T}\big[D(x\dot{\rightarrow}y) \rightarrow \big(T(x\dot{\rightarrow}y) \leftrightarrow (Tx \rightarrow Ty)\big)\big]$,

where $D(x)$ abbreviates $Tx \vee Fx$. Then, DT is DT^- plus full induction for $\mathcal{L}_T$.

As is desired, it is shown by straightforward meta-induction on $\phi(\vec{x})$ that

$$\mathsf{DT}^- \vdash \forall \vec{x}\big(D(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \rightarrow (T(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \leftrightarrow \phi(\vec{x}))\big).$$

DT-axioms look different from KF-axioms *prima facie* and they have distinct background motivations. However, DT and KF are in fact very closely related. In order to see this, we first consider the 'inner logic' behind DT.

In his consistency proof of DT, Feferman constructed a Kripkean fixed-point model for DT in which the truth of sentences are evaluated according to a special logic. This logic is what I called Feferman Logic in the last section, and I have presented the 'KF-like' system FKF there.

Then the next theorem indicates that determinate truth and iterative compositional truth bear a certain very close connection.

Theorem 2.7.2. DT (DT^-) and FKF+Cons (FKF^-+Cons) are identical systems.

Proof. The proof is straightforwardly done by showing all the axioms of one are provable in the other. We will only illustrate two typical examples. For $x, y \in \mathrm{St}_{\mathcal{L}_T}$, if $Tx \wedge Fx$ then Dx and thus $Fx \leftrightarrow \neg Tx$ in DT; thus $\mathsf{DT} \vdash$ Cons. For another example, in FKF+Cons, given any $x, y \in \mathrm{St}_{\mathcal{L}}$, if $T(x \dot{\rightarrow} y)$ then $(Tx \wedge Ty) \vee Fx$ by **FK9**. But it follows from Cons that $(Tx \wedge Ty) \vee \neg Tx$; we have $\neg Tx \vee Ty$. Conversely, if $\neg Tx \vee Ty$ and $D(x \dot{\rightarrow} y)$, then $\neg Tx \vee (Tx \wedge Ty)$ and Dx. Since $\neg Tx$ implies Fx by Dx, we have $(Tx \wedge Ty) \vee Fx$. Finally, $\mathsf{FKF} + \mathrm{Cons} \vdash$ **D11**. □

We can also relate WKF to certain 'DT-like' system of determinate truth. As was noted in fn.22, DT does not meet strong compositionality. Although there is a good reason for Feferman's restricting strong compositionality on the conditional '$\rightarrow$', it is then natural to consider a determinate system of truth with full strong compositionality. Such a system is obtained by replacing the axiom **D11** by the following:

$$\mathrm{St}_{\mathcal{L}_T}(x) \wedge \mathrm{St}_{\mathcal{L}_T} \rightarrow \big[D(x \dot{\rightarrow} y) \leftrightarrow (Dx \wedge Dy)\big].$$

Then, we can show in a parallel manner to the last theorem that the resulting system is identical with $\mathsf{WKF} + \mathrm{Cons}$.

Feferman [19] raised a conjecture about the proof-theoretic strength of DT.

Feferman's First Conjecture: $\mathsf{DT}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\varepsilon_0}$.

Feferman also suggested its proof. As far as I see, his suggested proof does work for the lower bound of $\mathsf{DT}[\![\mathsf{PA}]\!]$, but not for the upper bound as it is. He suggested to simulate the construction of Kripkean fixed-point model of $\mathsf{DT}[\![\mathsf{PA}]\!]$, which is given in his consistency proof of $\mathsf{DT}[\![\mathsf{PA}]\!]$, within

Σ^1_1-AC, as in the proof for the upper bound of KF$[\![$PA$]\!]$ in [17]. However, as we have seen, every fixed-point model of DT must be consistent in the sense that $Ta \wedge Fa$ for no a, and Σ^1_1-AC only assures the existence of a fixed-point of an Σ^1_1 positive operator but not necessarily consistent one. In the next chapter, we will give a solution to this obstacle and positive answer to the conjecture together with his second conjecture explained below as one application of relative truth definability.

2.8 Systems of Symmetric Truth

Systems like KF and DT are called type-free systems of truth, since they allow us to prove the truth of sentences which contain the truth predicate itself. However, any extension Q of KF or DT do not possess the following *symmetry* property of truth:

$$\mathsf{Q} \vdash \sigma \quad \text{if and only if} \quad \mathsf{Q} \vdash T\ulcorner\sigma\urcorner;$$

in my terminology of the last chapter, this symmetry property corresponds to the external T-schema (w.r.t. the default conditional at the meta-level); we will introduce the notions of 'inner theory' and 'outer theory' in Chapter 4, and the symmetry is equivalent to the coincidence the inner and outer theories of Q. As a corollary, neither of them can be consistently closed under Cons and Comp. In the present section, we present some systems of symmetric type-free truth.

We first introduce the system FS; the acronym FS comes from Friedman and Sheard who first introduced this system [23]. This appellation and the following formulation is due to Halbach [31].

Definition 2.8.1. The system FS over $\mathcal{L}_T$ is defined as B with full-induction for $\mathcal{L}_T$ plus the following axioms

FS1 $\forall \vec{x} \in \mathrm{CT} \to [T(\dot{R}\vec{x}) \leftrightarrow R(\mathrm{val}(\vec{x}))]$

FS2 $\forall x \in \mathrm{St}_{\mathcal{L}_T}[T(\dot{\neg}x) \leftrightarrow \neg Tx]$

FS3 $\forall x, y \in \mathrm{St}_{\mathcal{L}_T}\big[\big(T(x \dot{\vee} y) \leftrightarrow (Tx \vee Ty)\big) \wedge \big(T(x \dot{\rightarrow} y) \leftrightarrow (Tx \rightarrow Ty)\big)\big]$

FS4 $\forall z \forall x\big[\mathrm{St}_{\mathcal{L}_T}(\dot{\forall} z.x) \rightarrow \big(T(\dot{\forall} z.x) \leftrightarrow \forall y Tx(y/z)\big)\big]$,

and the next two extra inference rules: for each $\mathcal{L}_T$-sentence σ,

$$\frac{\sigma}{T(\ulcorner \sigma \urcorner)}\ (\mathrm{NEC}) \qquad \frac{T(\ulcorner \sigma \urcorner)}{\sigma}\ (\mathrm{CONEC})$$

Proposition 2.8.2. $\mathsf{FS} \vdash \mathrm{Cons} + \mathrm{Comp}$.

The proof-theoretic strength of FS was determined by Halbach [31]. FS seems natural and has desirable properties, i.e., Cons and Comp, but it turned out to be ω-inconsistent as can be shown by using McGee's trick [57].

Theorem 2.8.3. **(1)** $\mathsf{FS} \equiv \mathsf{RT}_{<\omega}$. **(2)** FS is ω-inconsistent.

Hence, when $\mathsf{B} = \mathsf{PA}$, we have $\mathsf{FS}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\omega}$. Rathjen [68] showed that $\mathsf{ACA}_0 + \mathrm{BR}$, where BR denotes the Bar Rule, and ACA_0^+, i.e., ACA_0 plus the axiom of the existence of ω-th Turing Jump of each set X,[23] prove the same Π^1_1 theorems. Then we can observe from Rathjen's proof that $\mathsf{RA}_{<\omega}$ (and thus $\mathsf{FS}[\![\mathsf{PA}]\!]$) also has the same arithmetical theorems as they have.[24]

We next look at another system of symmetric truth. Let us consider natural generalizations of (NEC) and (CONEC): the new rules NEC^* and CONEC^* respectively defined as

$$\frac{T^{\ulcorner}\phi(\dot{\vec{x}}^{\urcorner})}{\phi(\vec{x})}\ (\mathrm{NEC}^*) \qquad \frac{\phi(\vec{x})}{T^{\ulcorner}\phi(\dot{\vec{x}})^{\urcorner}}\ (\mathrm{CONEC}^*)$$

for each $\mathcal{L}_T$-formula $\phi(\vec{x})$. We notice that they are admissible in FS due to the presence of NEC

[23] For the precise definition and more explanations of ACA_0^+, see [77, Ch.X].

[24] It is noted that Rathjen, in fact, already showed that $\mathsf{RA}_{<\omega} \leq \mathsf{ACA}_0 + \mathrm{BR}$ $[\Pi^0_\infty]$ in [68]. The converse ($\mathsf{ACA}_0 + \mathrm{BR} \subset_{\Pi^0_\infty} \mathsf{RA}_{<\omega}$) is obtained by observing that the system ACA_0^ω, which Rathjen intermediately uses to establish his result, is embeddable in $\mathsf{RA}_{<\omega}$. Although the proof of the latter is model-theoretic, Theorem 3.5.6 below gives the proof-theoretical one, since $\mathsf{ACA}_0 + \mathrm{BR}$ can be embedded in $\mathsf{TC}^*[\![\mathsf{PA}]\!]$ ($\equiv \mathsf{RA}_{<\omega}$) in the standard manner (the second-order variables are interpreted to range over $\mathrm{St}_{\mathcal{L}_0(P)}$).

and CONEC (use FS4).

Let $\mathrm{Pos} \subset \mathcal{L}_T$ be the set of T-positive $\mathcal{L}_T$-formulae. There are some different but equivalent ways of defining T-positive formulae and in fact the subsequent arguments in the present chapter does not depend on the choice of its definition. However, for a technical reason, let us assume that Pos is defined by the following BNF:

$$\mathrm{Pos} ::= s(\vec{x}){=}t(\vec{x}) \mid Tt(\vec{x}) \mid \mathrm{Pos} \wedge \mathrm{Pos} \mid \mathrm{Pos} \vee \mathrm{Pos} \mid \forall z.\mathrm{Pos} \mid \exists z.\mathrm{Pos}$$

Then, POS is defined as $\mathrm{Pos} \cap \mathrm{St}_T$: i.e., PSt is the set of T-positive $\mathcal{L}_T$-sentences.

We will consider another KF-style system POSKF.

Definition 2.8.4. POSKF consists of PA, the full induction over $\mathcal{L}_T$ and:

P1 $x, y \in \mathrm{CT}(x) \to \big[T(x \dot{=} y) \leftrightarrow \mathrm{val}(x) = \mathrm{val}(y)\big] \wedge \big[T(x \dot{\neq} y) \leftrightarrow \mathrm{val}(x) \neq \mathrm{val}(y)\big]$

P2 $\mathrm{CT}(x) \to [T\mathrm{val}(x) \leftrightarrow T(\dot{T}x)]$

P3 $\mathrm{POS}_{\mathcal{L}}(x) \wedge \mathrm{POS}_{\mathcal{L}}(y) \to \big[T(x \dot{\wedge} y) \leftrightarrow (Tx \wedge Ty)\big]$

P4 $\mathrm{POS}_{\mathcal{L}}(x) \wedge \mathrm{POS}_{\mathcal{L}}(y) \to \big[T(x \dot{\vee} y) \leftrightarrow (Fx \vee Fy)\big]$

P5 $\mathrm{Var}(z) \wedge \mathrm{POS}(\dot{\forall} z.x) \to \Big[T(\dot{\forall} z.x) \leftrightarrow \forall y Tx(y/z)\big)\Big]$

P6 $\mathrm{Var}(z) \wedge \mathrm{POS}(\dot{\forall} z.x) \to \Big[T(\dot{\exists} z.x) \leftrightarrow \exists y Tx(y/z)\big)\Big]$,

where POS is the representation of the set POS of T-positive formulae.

Lemma 2.8.5. POSKF $\vdash$ PUTB; shown by meta-induction.

Hence, POSKF is a supersystem of PUTB. Indeed, POSKF is properly stronger than PUTB, since the former include $\mathcal{T}(\mathsf{PA})$ but the latter doesn't.

Our goal is to show the next theorem.

Theorem 2.8.6. $\mathsf{POSKF}+\mathrm{NEC}^*+\mathrm{CONEC}^*$ is conservative over POSKF for arithmetical sentences.

Then, since POSKF is truth-definable in PUTB, we have the next corollary:

Corollary 2.8.7. $\mathsf{PUTB}+\mathrm{NEC}^*+\mathrm{CONEC}^*$ is conservative over PUTB for arithmetical sentences.

Remark 6. $\mathsf{KF}+\mathrm{NEC}+\mathrm{CONEC}$ is inconsistent, and the same applies to $\mathsf{WKF}+\mathrm{NEC}+\mathrm{CONEC}$ and $\mathsf{DT}+\mathrm{NEC}+\mathrm{CONEC}$ as well. In addition, $\mathsf{PUTB} \subsetneq \mathsf{POSKF} \subsetneq \mathsf{KF}$ and $\mathsf{WKF} \not\subset \mathsf{POSKF}$ and $\mathsf{POSKF} \not\subset \mathsf{DT}$.

I will give the proof of Theorem 2.8.6 in the rest of the present section. It suffices to show that

> Given any derivation $\mathcal{D}$ from $\mathsf{POSKF}+\mathrm{NEC}^*+\mathrm{CONEC}^*$, for any $(M, D') \models \mathsf{POSKF}$, there exists $D \subset M$ such that $(M, D) \models \mathcal{D}$,

where $(M, D) \models \mathcal{D}$ means that the universal closure of each $\phi \in \mathcal{D}$ is true in (M, D); for, if $\mathsf{POSKF}+\mathrm{NEC}^*+\mathrm{CONEC}^* \vdash \sigma$ for some arithmetical σ, then σ is true in every $M \models \mathsf{PA}$ expandable to a model of POSKF, and thus $\mathsf{POSKF} \vdash \sigma$.

Take any $(M, D') \models \mathsf{POSKF}$. Let $\mathcal{D} = \langle \phi_0, \ldots, \phi_n \rangle$ be a derivation in POSKF in Hilbert style. We can assume without loss of generality that if ϕ_i is $Tt(\vec{x})$ then $Tt(\vec{x})$ is not repeated later: i.e., there is no $i < j \leq n$ such that $Tt(\vec{x}) \equiv \phi_j$. Then, we will construct a model (M, D) such that

$$(M, D) \models \mathsf{POSKF} + \forall \vec{x}\big(\phi_0(\vec{x}) \wedge \ldots \wedge \phi_n(\vec{x})\big).$$

For the sake of simplicity, we assume all the variables occurring in $\mathcal{D}$ is included in $\vec{x}$. Let $\theta_0(\vec{x}), \ldots, \theta_l(\vec{x})$ be the sentences such that θ_i $(i \leq n)$ is not T-positive, $T(\ulcorner \theta_i \urcorner) = \phi_j$ for some $j \leq n$ and $T(\ulcorner \theta_i(\dot{\vec{x}}) \urcorner)$ obtained by NEC^*.

Consider the following operation $\mathbf{P}\colon \mathcal{P}(M) \to \mathcal{P}(M)$:

$$\begin{aligned}
a \in \mathbf{P}(X) \Leftrightarrow\ & \exists b \exists c \big[\mathrm{CT}^M(b) \wedge \mathrm{CT}^M(c) \wedge a = b \dot{=} c \wedge \mathrm{val}^M(b) = \mathrm{val}^M(c)\big] \\
& \vee \exists b \exists c \big[\mathrm{CT}^M(b) \wedge \mathrm{CT}^M(c) \wedge a = b \dot{\neq} c \wedge \mathrm{val}^M(b) \neq \mathrm{val}^M(c)\big] \\
& \vee \exists \vec{d} \big[a = \mathrm{sb}^M_{\vec{x}}(\ulcorner \theta_0(\vec{x}) \urcorner, \vec{d}) \vee \ldots \vee a = \mathrm{sb}^M_{\vec{x}}(\ulcorner \theta_l(\vec{x}) \urcorner, \vec{d})\big] \\
& \vee \exists b \big[\mathrm{CT}^M(b) \wedge a = \dot{T}^M b \wedge \mathrm{val}(b) \in X\big] \\
& \vee \exists b \exists c \big[\mathrm{POS}^M(b) \wedge \mathrm{POS}^M(c) \wedge a = b \dot{\wedge}^M c \wedge b \in X \wedge c \in X\big] \\
& \vee \exists b \exists c \big[\mathrm{POS}^M(b) \wedge \mathrm{POS}^M(c) \wedge a = b \dot{\vee}^M c \wedge \big(b \in X \vee c \in X\big)\big] \\
& \vee \exists z \exists b \big[\mathrm{Var}^M(z) \wedge \mathrm{POS}^M(\dot{\forall} z.b) \wedge a = \dot{\forall}^M z.b \wedge \forall y[b(y) \in X]\big] \\
& \vee \exists z \exists b \big[\mathrm{Var}^M(z) \wedge \mathrm{POS}^M(\dot{\exists} z.b) \wedge a = \dot{\exists}^M z.b \wedge \exists y[b(y) \in X]\big],
\end{aligned}$$

where, for a predicate R or function f of $\mathcal{L}_{\mathsf{PA}}$, R^M is the interpretation of R or f in M.

Then, since $\mathbf{P}$ is a positive arithmetical operator and $\mathsf{POSKF}(\supset \mathsf{PUTB})$ can construct a fixed-point of such an operator (though not necessarily the least one) and (M, D') is a model of POSKF, there is $D \subset M$ which is definable in (M, D') and a fixed-point of $\mathbf{P}$.

More precisely, set $\xi(x, y)$ to be:

$$\begin{aligned}
&\exists b \exists c \big[\mathrm{CT}(b) \wedge \mathrm{CT}(c) \wedge x = b \dot{=} c \wedge \mathrm{val}(b) = \mathrm{val}(c)\big] \\
&\vee \exists b \exists c \big[\mathrm{CT}(b) \wedge \mathrm{CT}(c) \wedge x = b \dot{\neq} c \wedge \mathrm{val}(b) \neq \mathrm{val}(c)\big] \\
&\vee \exists \vec{d} \big[x = \mathrm{sb}_{\vec{x}}(\ulcorner \theta_0(\vec{x}) \urcorner, \vec{d}) \vee \ldots \vee x = \mathrm{sb}_{\vec{x}}(\ulcorner \theta_l(\vec{x}) \urcorner, \vec{d})\big] \\
&\vee \exists b \big[\mathrm{CT}(b) \wedge x = \dot{T} b \wedge Ty\big(\mathrm{val}(b)\big)\big] \\
&\vee \exists b \exists c \big[\mathrm{POS}(b) \wedge \mathrm{POS}(c) \wedge x = b \dot{\wedge} c \wedge Ty(b) \wedge Ty(c)\big] \\
&\vee \exists b \exists c \big[\mathrm{POS}(b) \wedge \mathrm{POS}(c) \wedge x = b \dot{\vee} c \wedge \big(Ty(b) \vee Ty(c)\big)\big] \\
&\vee \exists z \exists b \big[\mathrm{Var}(z) \wedge \mathrm{POS}(\dot{\forall} z.b) \wedge x = \dot{\forall} z.b \wedge \forall w[Ty(w)]\big] \\
&\vee \exists z \exists b \big[\mathrm{Var}(z) \wedge \mathrm{POS}(\dot{\exists} z.b) \wedge x = \dot{\exists} z.b \wedge \exists w[Ty(w)]\big],
\end{aligned}$$

Then diagonalize ξ and take its fixed-point formula γ such that

$$\gamma(x) \leftrightarrow \xi(x, \ulcorner \gamma \urcorner).$$

Note that since ξ is T-positive, we can assume that γ is also T-positive by the usual construction of fixed-point formulae. Then, we take $D := \gamma^M(x)$. We will use the same type of construction again in the next chapter.

Lemma 2.8.8. (M, D) is a model of POSKF.

Proof. First, PA-axioms are obviously satisfied since $M \models \mathsf{PA}$. Second, the full induction over $\mathcal{L}_T$ is trivially satisfied by (M, D), since D is definable in (M, D') and (M, D') is a model of the full induction. Finally, it suffices to show that P1-P6 are all satisfied by (M, D); but it can be straightforwardly shown by using the fixed-pointness of D. The important point is that, since $\theta_0, \ldots, \theta_l$ are not T-positive, they have nothing to do with the POSKF-axioms and thus do not

affect the usual proof. I will only illustrate one example: $(M, D) \models$ P3. Suppose $a = b \dot{\wedge}^M c$ and $b, c \in \mathrm{POS}^M$.

$$\begin{aligned}
\gamma^M(b \dot{\wedge}^M c) &\Leftrightarrow \xi(b \dot{\wedge}^M c, \ulcorner \gamma \urcorner) && \text{(Since } \gamma \text{ is a fixed-point of } \xi\text{)} \\
&\Leftrightarrow T^M\gamma(b) \wedge T^M\gamma(c) \\
&\Leftrightarrow \gamma^M(b) \wedge \gamma^M(c) && \text{(By PUTB)}
\end{aligned}$$ □

Now, we show that $(M, D) \models \mathcal{D}$ by induction on $\phi_k(\vec{x})$ $(k \leq n)$. By the last lemma, it suffices to show that $(M, D) \models \phi_k(\vec{c})$ for all $\vec{c} \in M$ if ϕ_k is obtained by NEC* or CONEC*.

For the former case, let $\phi_k \equiv T\big(\ulcorner \psi(\dot{\vec{x}}) \urcorner\big)$ obtained by NEC from $\psi(\vec{x})$. Suppose $\psi(\vec{x})$ is T-positive. Then since (M, D) is a model of POSKF and we already have $(M, D) \models \psi(\vec{c})$ for all $\vec{c}$ by IH, it follows from PUTB-schema (recall that POSKF $\vdash$ PUTB) that $(M, D) \models \phi_k(\vec{c})$ for all $\vec{c}$. Otherwise, $\psi(\vec{x})$ is θ_i for some $i \leq l$, and thus, for any $\vec{c} \in M$, $\phi_k(\vec{c}) \in D$ by the third line of the definition of $\mathbf{P}$.

Finally, suppose $\phi_k(\vec{x})$ is obtained by CONEC* from $\phi_j(\vec{x}) \equiv T\big(\ulcorner \phi_k(\dot{\vec{x}}) \urcorner\big)$ $(j < k)$. If ϕ_k is T-positive, it is satisfied by (M, D) again since it is a model of PUTB; Suppose not. By IH, we have $(M, D) \models \phi_j(\vec{c})$ for all $\vec{c} \in M$ and thus $\ulcorner \phi_k(\vec{c}) \urcorner \in D$ for all $\vec{c}$. Since ϕ_k is not T-positive and thus $\ulcorner \phi(\vec{c}) \urcorner \notin \mathrm{POS}^M$ for all $\vec{c}$ (see the above remark), ϕ_k is identical to one of $\theta_0, \ldots, \theta_l$; for, if $\ulcorner \phi_k(\vec{c}) \urcorner \in D$ then it must meet one of the disjuncts of the defining clauses of $\mathbf{P}$; but then, only the second line of the definition of $\mathbf{P}$ does apply. Let $\phi_k = \theta_i$. Then, by definition, $T(\ulcorner \theta_i \urcorner)$ is obtained by NEC* from some ϕ_m $(m \leq n)$. However, $T(\ulcorner \theta_i \urcorner)$ is just identical to ϕ_j and we have already assumed that Tt is not repeated in $\mathcal{D}$; therefore, $m < j$ and thus by IH $(M, D) \models \phi_m(\vec{c})$ for all $\vec{c} \in M$; i.e., $(M, D) \models \phi_k(\vec{c})$.

We have completed the proof.

Indeed, the above proof can be directly transformed into proof-theoretic proof; given a proof

$\mathcal{D}$, then we just replace each T by a fixed-point of $\mathbf{P}$ which is effectively defined in POSKF, since $\mathsf{POSKF} \supset \mathsf{PUTB} \supset \mathrm{I}\hat{\mathrm{D}}_1$.

Furthermore, we can directly apply the above argument to POSKF^- and PUTB^- as it is. Now we have the next corollary.

Corollary 2.8.9. $\mathsf{POSKF}^- + \mathrm{NEC}^* + \mathrm{CONEC}^*$ is conservative over POSKF for arithmetical sentences, and the same applies to PUTB^- as well.

2.9 Systems of iterative non-compositional truth

Cantini [6] introduced the system VF of self-applicable truth, which is designed so as to embody the Kripkean fixed-point semantics with van Fraassen's supervaluation schema, and showed that it is proof-theoretically equivalent to the system ID_1 of elementary inductive definitions.

Given a system S, Ax_{S} represents the set of the axioms of S and $\mathrm{Prv}_{\mathsf{S}}$ is a canonical provability predicate of S. Let $\mathsf{PL}_{\mathcal{L}}$ be pure classical predicate logic for the language $\mathcal{L}$; then $\mathrm{Ax}_{\mathsf{PL}_{\mathcal{L}_T}}$ represents the set of logical axioms (for the language $\mathcal{L}_T$). I adopt the simpler version of VF in [7] (or [45]) rather than the original one in [6].

Definition 2.9.1. The system VF^- over $\mathcal{L}_T$ comprises PA plus the following truth axioms:

V1 $\forall \vec{x}[T(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \to \phi(\vec{x})]$, for each $\mathcal{L}_T$-formula $\phi(\vec{x})$.

V2 $\forall \vec{x} \in \mathrm{CT}[(T(\dot{R}\vec{x}) \leftrightarrow R\vec{x}^\circ) \wedge (F(\dot{R}\vec{x}) \leftrightarrow \neg R\vec{x}^\circ)]$, for each $R \in \mathcal{L}_0$

V3 $\forall x \in \mathrm{St}_{\mathrm{T}}[\mathrm{Ax}_{\mathrm{PL}_{\mathcal{L}_T}}(x) \to Tx]$

V4 $\forall x, y \in \mathrm{St}_{\mathrm{T}}[T(x \dot{\to} y) \to (Tx \to Ty)]$

V5 $\forall z \forall x[\mathrm{St}_{\mathrm{T}}(\dot{\forall} z.x) \to (\forall y Tx(y/z) \to T\dot{\forall} z.x)]$

Then, the system VF is obtained by adding full-induction for $\mathcal{L}_T$ to VF^-.

Theorem 2.9.2 (Cantini [6, 7])**.** $\mathsf{VF}^- \equiv \mathsf{B}$ and $\mathsf{VF} \equiv \mathsf{ID}_1$.

We will later have a closer look at iterative non-compositional truths in Chapter 5.

2.10 Schematic Reflective Closures

Feferman [17] introduced the notion of *schematic reflective closure* as an answer to the question 'which schemata ... in the language of [a theory] ought to be accepted if one has accepted the basic (schematic) axioms and rules of [the theory] ([17, p.2])'.[25]

Definition 2.10.1. Given a base system B over $\mathcal{L}_0$, let $\mathsf{B}(P)$ denote a system over $\mathcal{L}_0(P) := \mathcal{L}_0 \cup \{P\}$, where P is a new unary predicate symbol, whose non-logical axioms are just the same as B (i.e., P is a vacuous predicate in $\mathsf{B}(P)$ for which only logical axioms are assumed).

Then, the schematic reflective closure Q^* of a system Q of truth with a base system B, is a system over $\mathcal{L}_{\mathsf{Q}}(P) = \mathcal{L}_{\mathsf{Q}} \cup \{P\}$ whose axioms consist of $\mathsf{Q}[\![\mathsf{B}(P)]\!]$, full induction schema for $\mathcal{L}_{\mathsf{Q}}(P)$, and a new inference rule P-Subst:

$$\frac{\phi(P)}{\phi\big(\hat{x}\psi(x)\big)}\ (P\text{-Subst}), \tag{2.3}$$

for any formula $\phi(P)$ of $\mathcal{L}_0(P)$ (possibly interspersed with occurrences of P) and a formula ψ of $\mathcal{L}_{\mathsf{Q}}(P)$, where $\phi\big(\hat{x}\psi(x)\big)$ is the result of replacing each occurrence of Pt in ϕ by $\psi[t/x]$ and untangling bound variables.[26] Following Feferman's notation, we sometimes write $\phi(\hat{\psi})$ for $\phi(\hat{x}\psi(x))$.

Notice that $\mathsf{Q}[\![\mathsf{B}(P)]\!]$ contains truth-axioms for the new predicate symbol P. For example,

[25] The definition below is modified for our current arguments and not exactly the same as Feferman's original one. Feferman's original definition depends on what is regarded as 'schematic axiom'. At any rate, if we focus on arithmetical base theories and only regard induction schema as 'schematic axiom', then Feferman's definition and ours coincide.

[26] As Feferman pointed out, we should not allow arbitrary ϕ of $\mathcal{L}_{\mathsf{Q}}(P)$ in the premise of P-Subst; otherwise, it may leads to a contradiction. For example, if we allow arbitrary ϕ in KF^*, then we obtain $T(\dot{P}x) \leftrightarrow \psi$ for all ψ.

$\mathsf{KF}[\![\mathsf{B}(P)]\!]$ contains the axiom

$$\mathrm{CT}(x) \to \big[P\big(\mathrm{val}(x)\big) \leftrightarrow T(\dot{P}x)\big],$$

and each '$\mathrm{St}_{\mathcal{L}_T}$' occurring in **K0**-**K8** is replaced by $\mathrm{St}_{\mathcal{L}_T(P)}$.

Schematic reflective closure can be viewed as an analogue to the Bar rule for sybsystems of second-order arithmetic (cf. [17, §5]). It is not only a philosophically significant framework but also a useful and natural device of extending a given theory. It indeed increases the proof-theoretic strength of theories in many cases (though we will see some exceptional cases in §5).

Theorem 2.10.2 (Feferman [17])**.** $\mathsf{KF}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$. Thus $\mathsf{KF}^*[\![\mathsf{PA}]\!]$ is also proof-theoretically equivalent to, e.g., Σ^1_1-$AC + \mathrm{BR}$, $\widehat{\mathsf{ID}}_{<\omega}$ and ATR_0.[27]

In addition to his first one, Feferman also conjectured that

Feferman's Second Conjecture: $\mathsf{DT}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$.

We will later show that this is indeed true.

We have now introduced all the systems to be considered in the rest of the present thesis. However, before concluding this section, we need to introduce some extra systems which are to be intermediately used for the proof-theoretic analyses of schematic reflective closures.

[27] For the definition of $\widehat{\mathsf{ID}}_{<\omega}$, see [14]. For ATR_0, see [77].

Chapter 3

Relative Truth Definability

In the present chapter, I will propose one new tool, *relative truth definability*, to compare more conceptual aspects of systems of truth. We focus on the cases in which systems of truth share their base system. This is because we are interested in a comparison of the conceptions of truth and thus we want to keep all the other conceptions fixed thereby excluding inessential factors. For example, suppose the base system of one system is PA and that of another system is ZFC. Then, in most cases, the truth of the former system can be defined in the latter solely in terms of sets without appealing to its truth predicate. In such cases, what conception of the truth is employed in the latter system does not matter at all. Hence, we start by fixing a base system B to which every system is obtained by adding truth predicate(s) and its (their) axioms and inference rules.[1]

3.1 Comparing Conceptual Aspects of Axiomatic Truth

We have seen a large variety of axiomatic systems of truth. Different systems embody different views and conceptions of truth behind their axioms. It should be thus expected that one can gain

[1]The contents of the present section are based on my [24].

both mathematically and philosophically valuable insights about truth by investigating how they are related to each other and by comparing them from various points of view.

In the last chapter, we have also seen that much proof-theoretic work on axiomatic systems of truth has been done. Various systems of truth have been related to other systems of truth or certain subsystems of second-order arithmetic in terms of proof-theoretic equivalence, equiconsistency. For example, we have seen that some systems of truth have the same truth-free theorems, even if they apparently embody different views of truth. In many cases, these proof-theoretic works focus only on the arithmetical parts of systems, i.e., their theorems containing no truth predicate. The coincidences of truth-free theorems might indeed result from a certain close and intimate conceptual connection between two systems, but it seems that mere coincidences of truth-free theorems cannot solely imply the existence of such a connection. For example, let T be a system over $\mathcal{L}_T$ defined as B plus a single additional axiom '$T(\ulcorner 0 = 0 \urcorner)$'. Then, we can show that T and UTB are proof-theoretically equivalent (and in fact mutually interpretable if B is reflexive). However, it seems implausible to think that T and UTB share a common view and conception of truth, or, in other words, it is hardly agreed that they are conceptually 'equivalent'. Hence, proof-theoretic equivalence, conservativity or other well-known intertheoretic relations seem too coarse-grained for our purpose of comparing conceptual aspects of systems of truth.

If one adopts an instrumentalist approach to truth, in which truth is regarded merely as a practical device to increase the deductive or expressive power of systems, it might be sufficient to consider only what kind of and how rich truth-free consequences can be obtained by means of a truth predicate (or truth predicate*s* if the language has more than one truth predicates). However, if we are interested in more conceptual aspects of truth systems, it seems no longer sufficient to consider only the derivability of truth-free sentences. For, different systems based on different and incompatible views of truth may have the same truth-free consequences just by accident. Hence, in order to compare conceptual aspects of systems of truth, we must take into consideration their

theorems which may contain truth predicate(s). We need mathematically finer-grained means to compare and distinguish their conceptual aspects.

We are thus interested in a more fine-grained means of comparing conceptual aspects of systems of truth. In the present section, we will propose such a criterion, by means of which we can represent a certain 'equivalence' or 'reducibility' among systems of truth from a more conceptual point of view, although it may be still too coarse to fully represent 'conceptual equivalence' or 'conceptual reducibility' (in some strong sense). As we have argued, any relation among systems which only focuses on truth-free parts seems insufficient for our purpose on the one hand. On the other hand, some relations are too strong. For example, subsystem relation and its induced equivalence relation, i.e., identity (as systems), should be too strong; e.g., $\mathsf{KF} + \neg T(\mathbf{n})$ for some numeral $\mathbf{n}$ which codes no sentence is not a subsystem of KF, but we usually think that the difference only consists in an inessential point irrelevant to their views of truth.[2]

In the present and next chapter, I will try to suggest two alternative means more suitable for comparing conceptual aspects of systems of truth than other so far presented means. Namely, what we will suggest is *relative truth definability*, which examine whether the truth predicate(s) of one system Q is definable in another system S in terms of S's truth predicate(s). Although its precise definition and detailed explanations will be given in §3, I briefly explain it here. Given two systems Q and S of truth over languages $\mathcal{L}_\mathsf{Q}$ and $\mathcal{L}_\mathsf{S}$ which share the same base system B over $\mathcal{L}_\mathsf{B}$ ($\subset \mathcal{L}_\mathsf{Q} \cap \mathcal{L}_\mathsf{S}$), we say that Q is relatively truth definable in S, if for each truth predicate T of $\mathcal{L}_\mathsf{Q}$ there exists a predicate R_T of $\mathcal{L}_\mathsf{S}$ such that $\mathsf{Q} \vdash \phi$ implies $\mathsf{S} \vdash \phi'$, where ϕ' is obtained from ϕ by replacing all the occurrence of each truth predicate T by R_T; in other words, Q is relatively truth definable in S, if there exists $\mathcal{L}_\mathsf{B}$-conservative interpretation of Q in S. This notion of relative truth definability is in fact a finer-grained tool for comparing systems of truth than mere derivability of truth-free

[2]Subsystem relation as a representation of intertheoretic 'reducibility' was also criticized by Niebergall [63]. His argument is about 'system reducibility' in general and thus is applied to our current discussion as well.

sentences or other well-known means. For example, conservativity for truth-free theorems and relative interpretability follow from relative truth definability, but the converse does not necessarily hold. We will then argue that relative truth definability sheds more light on conceptual aspects of systems of truth and compare various systems in the light of it. In the end, we will see that relative truth definability can distinguish some systems with the same truth-free theorems but based on apparently different views of truth and also that it is useful in its applications.

I do not at all intend to deny other kinds of comparison of axiomatic systems of truth. What the present chapter aims at is rather to suggest one possible methodology for comparison and investigation of more conceptual aspects of truth systems, to which not enough attention has been paid, and to give a new perspective of recent works of axiomatic truth systems in the light of relative truth definability.

One probably comes to relative interpretability as a candidate for such an intertheoretic relation. Indeed, we often say, for instance, that geometry (or any other subject of mathematics) is reducible to set theory, and this usually means the existence of a relative interpretation of geometry in set theory. However, it seems still arguable whether this implies that the concept of geometrical objects such as 'line' or 'triangle' are 'reducible' to the concept of set. It is a deep and difficult problem whether this is a genuine reduction or not, but, at any rate, relative interpretability is not what we want to propose.[3]

First of all, we note that the system T above and UTB are mutually interpretable. Hence relative interpretability cannot exclude this kind of pathological cases. In general, the problems of relative interpretability as our means of comparison are:

(A) Relative interpretation may interpret terms and predicates by totally different notions;

(B) Relative interpretation may change the range of quantifiers, in other words, the ontological

[3]For more discussion of relative interpretability as a tool of reduction in more general settings, particularly on its role in foundational issues, see e.g., [18] and [63].

commitment of the interpreted system;

(C) Relative interpretability does not necessarily preserve truth-free theorems;

(D) Relative interpretability just follows from conservativity for truth-free theorems in some typical setting of systems of truth.

We will explain each of these problems below.

The characteristics (A) and (B) of relative interpretability are principal defects for our current objectives. It is true that relative interpretability is very useful and widely applicable for comparing diverse kinds of mathematical systems, but we are currently in a special situation in which only systems of truth are considered and we compare only their conceptions of truth. For example, relative interpretability is often regarded as a suitable tool for ontological reduction of mathematical (or other) objects posited by one mathematical subject into those posited by another subject. However, what we are aiming at is not ontological reduction but rather, so to say, 'conceptual reduction'. Thus, for our current purpose, we would like to keep every other concept fixed in order to exclude inessential factors, although it may make our approach less versatile than relative interpretability. Therefore, since relative interpretability may require that the words, e.g., 'for all natural numbers' or 'greater than' of one system be translated by 'for all even numbers' or 'less than' in another system, relative interpretability seems unsuitable for our current purpose.

The problem (C) is caused by (A) and (B) and it is problematic since we are looking for a finer-grained means to compare systems. There are indeed two systems Q_0 and Q_1 such that Q_1 can prove properly more truth-free theorems than Q_0 but is relatively interpretable in Q_0; e.g., consider TC and $\mathsf{TC} + \neg\mathrm{Cons}(\mathsf{TC})$. If the truth of Q_1 is 'reducible' to Q_0, then it is natural to require that the truth of Q_1 should not bring richer truth-free consequences than that of Q_0 does.

The fourth problem (D) is more precisely elaborated by the following fact.

Theorem 3.1.1. Let Q and S be first-order classical systems and primitive recursive extensions of

PA. Suppose that Q and S are both reflexive and that $\mathsf{Q} \subset_{\Pi^0_1} \mathsf{S}$. Then Q is interpretable in S.

Proof. Let $\sigma(x)$ be a primitive recursive predicate which binumerates Q-axioms. Since Q is reflexive, $\mathsf{Q} \vdash \mathrm{Con}_{\sigma|k}$ and thus $\mathsf{S} \vdash \mathrm{Con}_{\sigma|k}$ for each $k \in \mathbb{N}$. Set $\sigma^* := \sigma(x) \wedge \mathrm{Con}_{\sigma|x}$. Then, σ^* binumerates Q in S. Finally, the claim follows from the arithmetized completeness theorem.[4] □

Many of the systems of truth are reflexive; in fact, as we shall see in §4, all the systems of truth we have so far introduced are reflexive. Hence, in many cases, only the truth-free part (indeed Π^0_1 part) of systems does matter to relative interpretability, and relative interpretability is no more fine-grained than the derivability of truth-free theorems.[5]

3.2 Relative Truth Definability

Now we present our own proposal, relative truth definability. Informally speaking, that a system Q is truth-definable in another system S means that the language $\mathcal{L}_\mathsf{Q}$ of Q can be translated into the language $\mathcal{L}_\mathsf{S}$ of S in such a way that the basic terms are kept unchanged and the concept of truth captured in Q is expressible and its function in Q can be fully simulated by a single formula in $\mathcal{L}_\mathsf{S}$ which translated the truth predicate(s) of $\mathcal{L}_\mathsf{Q}$. The formal and precise definition is the following.

Definition 3.2.1. Let Q and S be (primitive recursive) systems of truth over languages $\mathcal{L}_\mathsf{Q}$ and $\mathcal{L}_\mathsf{S}$ respectively, and let $\mathcal{L}_\mathsf{Q}$ be $\mathcal{L}_0 \cup \{T_i\}_{i \in I}$ where I is a certain index set; recall that $\mathcal{L}_\mathsf{Q}$ may have more than one truth predicates as in the case of $\mathsf{RT}_{<\alpha}$. In order to keep the definition as general as possible, the base systems B of Q and B' of S may be different with the proviso that $\mathcal{L}_\mathsf{B} \subset \mathcal{L}_{\mathsf{B}'}$, although we usually consider the cases in which they are the same.

Given a formula θ_i of $\mathcal{L}_\mathsf{S}$ for each $i \in I$, we define a translation $\mathcal{T}_{\vec{\theta}}$ from $\mathcal{L}_\mathsf{Q}$ to $\mathcal{L}_\mathsf{S}$ in the following way:

[4] See §6 and §8 of [54].

[5] Some non-reflexive systems of truth are known. For example, Cantini's $\mathsf{KF}_{\mathrm{int}}$ and its variant $\mathsf{KF}_{\mathrm{tot}}$ are not reflexive, since they are finitely axiomatizable.

$$\mathcal{T}_{\vec{\theta}}(\phi) := \begin{cases} \phi & \text{if } \phi \text{ is an atomic formula of } \mathcal{L}_0 \\ \theta_i(x) & \text{if } \phi \text{ is of the form } T_i(x) \\ \neg\mathcal{T}_{\vec{\theta}}(\psi) & \text{if } \phi \text{ is of the form } \neg\psi \\ \mathcal{T}_{\vec{\theta}}(\psi_0) \vee \mathcal{T}_{\vec{\theta}}(\psi_1) & \text{if } \phi \text{ is of the form } \psi_0 \vee \psi_1 \\ \forall x\mathcal{T}_{\vec{\theta}}(\psi) & \text{if } \phi \text{ is of the form } \forall x\psi \end{cases}$$

We say Q is *relatively truth definable* (more simply *truth-definable*) in S, or S (relatively) *defines the truth of* Q, when there exists formulae $\theta_i(x)$ of $\mathcal{L}_\mathsf{S}$ for each $i \in I$ such that $\mathsf{Q} \vdash \phi \Rightarrow \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi)$ for all $\phi \in \mathcal{L}_\mathsf{Q}$. Then, we say that $\mathcal{T}_{\vec{\theta}}$ is a *relative truth definition* (or simply *truth-definition*) of Q in S and we call θ_i *truth-defining formula* of the truth predicate T_i of Q in S.

In other words, Q is relatively truth definable in S iff there exists an $\mathcal{L}_0$-conservative relative interpretation of Q in S, or, equivalently, iff Q is a subsystem of a definitional expansion of S (with the proviso that the truth predicates of $\mathcal{L}_\mathsf{Q}$ and $\mathcal{L}_\mathsf{S}$ are syntactically distinguished in the definitional expansion). In the literature, this kind of interpretation has been occasionally used for determining proof-theoretic strength of systems and the idea and technique themselves are nothing new; e.g., Feferman gave an $\mathcal{L}_\mathsf{PA}$-conservative interpretation of $\mathsf{KF}[\![\mathsf{PA}]\!]$ in Σ^1_1-AC to establish $\mathsf{KF}[\![\mathsf{PA}]\!] \leq \Sigma^1_1$-$\mathsf{AC}$.

We will explain below that relative truth definability is finer-grained than other intertheoretic relations. First of all, we note that relative truth definability is at least as fine-grained as relative interpretability, consistency strength and conservativity of truth-free theorems: a relative truth definition is a relative interpretation; if a system Q is relatively truth definable in S then the truth-free theorems of the former are included in the latter, i.e., $\mathsf{Q} \subset_{\mathcal{L}_0} \mathsf{S}$, and thus $\mathsf{Q} \leq_{\mathrm{Con}} \mathsf{S}$.

On the other hand, we cannot generally obtain proof-theoretic reducibility from relative truth definability, since the latter has no condition on the system in which the definability itself is proved, although in practice, at least in all the cases in the present thesis, a proof of relative truth definability immediately entails proof-theoretic reducibility.

According to Feferman's original formulation in [16], a system Q is proof-theoretically reducible

to S for a set of formulae $\Phi \subset \mathcal{L}_{\mathsf{Q}} \cap \mathcal{L}_{\mathsf{S}}$, if there exists a partial recursive function f such that

(I) for each proof p in Q of a formula $\phi \in \Phi$, $f(p)\downarrow$ and is a proof in S of ϕ

(II) the formalization of (I) is provable in S.

This definition itself is indeed debatable, since, as was pointed out by Niebergall [63], it forfeits transitivity of proof-theoretic reducibility. However, under a moderate assumption, this problem is resolved. Let us say that Q is proof-theoretically reducible to S for Φ *provably in* a system T when (I) holds for Q and S and, instead of (II), (I) is provable in T. Then, the following fact due to Niebergall and Rathjen is known[6]:

Theorem 3.2.2. For any (primitive recursive) extensions Q and S of $\mathsf{I}\Sigma_1$, if Q is proof-theoretically reducible to S for a primitive recursive set Φ provably in S and the reduction function f can be taken to be primitive recursive, then Q is proof-theoretically reducible to S for Φ provably in $\mathsf{I}\Sigma_1$.

In practice, the reduction function f has been always primitive recursive in the literature: indeed, Rathjen rather defines proof-theoretic reducibility in [69] with the restriction that f be primitive recursive. In addition, we usually deal with systems stronger than $\mathsf{I}\Sigma_1$ in the context of foundational discussions and the present paper only considers such systems. Thus we can adopt, as our definition, proof-theoretic reducibility provably in $\mathsf{I}\Sigma_1$ via a primitive recursive reducing function, instead of Feferman's original definition.

Then, if we want relative truth definability to generally imply proof-theoretic reducibility, we need to add some new condition to Definition 3.2.1. For example, we can add the following condition (♭) to Definition 3.2.1:

(♭) The θ_i's are primitive recursively specified for each truth predicate T_i of Q, and it is provable

[6]The proof of Theorem 3.2.2 appeared in [18], which Feferman accredits to Niebergall, whereas Rathjen also pointed out the same fact in [69, §2.5].

in $\mathsf{I\Sigma_1}$ that if $\mathsf{Q} \vdash \phi$ then $\mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi)$; more formally,

$$\mathsf{I\Sigma_1} \vdash \forall x \in \mathrm{St}_{\mathsf{Q}}\big[\mathrm{Prv}_{\mathsf{Q}}(x) \to \mathrm{Prv}_{\mathsf{S}}\big(\mathcal{T}_{\vec{\theta}}(x)\big)\big], \tag{3.1}$$

where $\mathrm{Prv}_{\mathsf{Q}}$ ($\mathrm{Prv}_{\mathsf{S}}$) are canonical representations of provability in Q (S).

Then, the definition of truth-definability augmented by ($\flat$) implies proof-theoretic reducibility in general, since we can obtain by a well-known metamathematical property of $\mathsf{I\Sigma_1}$ that (3.1) is equivalent to

$$\mathsf{S} \vdash \forall x \in \mathrm{St}_{\mathsf{Q}} \forall y\big[\mathrm{Proof}_{\mathsf{Q}}(y, x) \to \mathrm{Proof}_{\mathsf{S}}\big(g(y), \mathcal{T}_{\vec{\theta}}(x)\big)\big],$$

for some primitive recursive g, where $\mathrm{Proof}_{\mathsf{Q}}(y, x)$ ($\mathrm{Proof}_{\mathsf{S}}(y, x)$) canonically expresses that y is a code of a proof in Q (S) of the formula coded by x. Furthermore, if Q is a first-order classical system without extra inference rules, we only need: e.g.,

($\flat'$) The θ_i's are primitive recursively specified for each T_i of Q, and we can construct a proof of $\mathcal{T}_{\vec{\theta}}(\sigma)$ for each Q-axiom σ primitive recursively on σ.

I think that these additional conditions neither lessen applicability and usability of truth-definability in practice nor harm the philosophical significance as a tool for comparing conceptual aspects of truth systems, although more careful discussions and philosophical evaluations are left for another occasion. At any rate, if one does not stick to general implication from relative truth definability to proof-theoretic reducibility, she need not worry too much about the issue argued so far, since relative truth definability immediately implies proof-theoretic reducibility in practice. In fact, in the subsequent proofs of relative truth definability in §4, we first primitive recursively assign a formula θ of $\mathcal{L}_{\mathsf{S}}$ to each truth predicate T of $\mathcal{L}_{\mathsf{Q}}$ and then explicitly prove that $\mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\sigma)$ for all Q-axioms σ and $\mathcal{T}_{\vec{\theta}}$ preserves each inference rule of Q. Given such a translation $\mathcal{T}_{\vec{\theta}}$, we can primitive recursively transform any proof of an $\mathcal{L}_0$-sentence in Q to a proof of the same sentence in

S.

Whether we add some additional conditions like ($\flat$) to its definition or not, relative truth definability can be proved to be transitive; note that it is trivially reflexive. Hence relative truth definability induces a pre-ordering among systems of truth; then the classes of mutually truth-definable systems form equivalent classes. Niebergall [63] raised five conditions for a good axiomatization of the reducibility relation ρ: for classical first-order systems Q and S,

PRL1 $\mathsf{Q} \subset \mathsf{S} \Rightarrow \mathsf{Q}\rho\mathsf{S}$;

PRL2 ρ is transitive;

PRL3 if $\mathsf{Q}\rho\mathsf{S}$ then Con(S) implies Con(Q);

PRL4 for each finite $\mathsf{Q}' \subset \mathsf{Q}$, if $\mathsf{Q}\rho\mathsf{S}$, there is a finite $\mathsf{S}' \subset \mathsf{S}$ such that $\mathsf{Q}'\rho\mathsf{S}'$;

PRL5 $\mathsf{I}\Sigma_1 \vdash \mathrm{Con}(\mathsf{S}) \to \mathrm{Con}(\mathsf{Q})$ for finitely axiomatized Q and S with $\mathsf{Q}\rho\mathsf{S}$.

It is easily shown that relative truth definability meets them all.

I would like to argue below that relative truth definability is a more suitable means than relative interpretability for expressing 'conceptual reducibility' of one system to another. First, the problems (A)-(D) (in §1) of relative interpretability no longer apply to relative truth definability. We have already observed that (C) is not the case for truth-definability; (D) doesn't hold for relative truth definability as we shall see in §4; since the base language and quantification are kept unchanged in relative truth definition, neither (A) nor (B) is the case. Second, some pathological examples like T (i.e., $\mathsf{B} + T(\ulcorner 0{=}0 \urcorner)$) above and UTB are excluded by relative truth definability.

Proposition 3.2.3. UTB is not truth-definable in T.

Proof. Let T' be T plus $\forall x(x \neq \ulcorner 0 = 0 \urcorner \to \neg Tx)$. Suppose UTB is truth-definable in T and thus T'. Note that $\mathsf{T}' = \mathsf{B} + \forall x[Tx \leftrightarrow x = \ulcorner 0{=}0 \urcorner]$. Then, if $\mathsf{T}' \vdash \phi$, the formula ϕ' obtained from ϕ by replacing each occurrence of $T(t)$ by $t = \ulcorner 0 = 0 \urcorner$ is derivable in B. This implies that B can define the truth of UTB: this contradicts Tarski's undefinability theorem. □

3.2.1 Mathematical Properties of Truth-Definability

We have mainly looked at the philosophical significance of relative truth definability, but it is also mathematically and technically useful. In the present subsection, we will see some of its mathematical properties and applicabilities.

First of all, as we have already observed, for systems Q and S, if Q is truth-definable in S, then $\mathsf{Q} \subset_{\mathcal{L}_0} \mathsf{S}$, Q is interpretable in S, $\mathsf{Q} \leq_{\mathrm{Con}} \mathsf{S}$, and $\mathsf{Q} \leq \mathsf{S}$ under some moderate condition like ($\flat$).

The next proposition immediately follows from the definition.

Proposition 3.2.4. **(1)** Suppose Q is truth-definable in S and $M \models \mathsf{B}$ can be expanded to a model of S (i.e., there exist $X_i \subset |M|$ for each truth predicate T_i of S such that $(M, \langle X_i \rangle_{i \in I}) \models \mathsf{S}$). Then M is also expandable to a model of Q.

(2) Suppose Q is truth-definable in S. If Q is ω-inconsistent, so is S.

(3) Suppose Q is truth-definable in S. Then Q plus full induction for $\mathcal{L}_{\mathsf{Q}}$ is truth-definable in S plus full induction for $\mathcal{L}_{\mathsf{S}}$.

Thus, relative truth definability preserves useful mathematical properties. Proposition 3.2.4 tells us that relative truth definability can distinguish pathological systems like ω-inconsistent system FS from others, since no ω-consistent system can define the truth of an ω-inconsistent system. It also tells that relative truth definability implies 'conservativity in the semantic sense for $\mathcal{L}_0$'; here a system S is conservative in the semantic sense over Q for $\mathcal{L}_0$ iff the $\mathcal{L}_0$-reduct of every model of Q is expandable to a model of S. McGee [60] argued that conservativity in the semantic sense is to be preferred to and has more philosophical significance than conservativity in the ordinary sense ('conservativity in the proof-theoretic sense' in the terminology of [60]), since, he claims, there is no metaphysical cost in moving from one system of truth to another conservative system in the semantic sense and this move between conservative systems in the semantic sense makes no difference about what the world is like. On the other hand, for instance, $\mathsf{RT}_{<\omega}$ and FS are conservative over each

other for $\mathcal{L}_0$ in the ordinary sense but the latter makes a stronger ontological commitment than the former since FS is ω-inconsistent.

In contrast to relative truth definability, relative interpretability and proof-theoretic reducibility satisfy none of (1)-(3) above. It is thus observed that relative truth definability is a finer-grained tool for comparing systems of truth than relative interpretability and proof-theoretical reducibility (and thus consistency strength and conservativity).

We will show one more important feature of relative truth definability in its application to Feferman' *schematic reflective closure* already explained in Ch.2, §10. The next lemma shows that truth-definability carries over to schematic reflective closures.

Lemma 3.2.5. Let Q be a first-order classical system of truth with a base system B (without any extra inference rules such as NEC and CONEC). Suppose $\mathsf{Q}[\![\mathsf{B}(P)]\!]$ is truth definable in $\mathsf{S}[\![\mathsf{B}(P)]\!]$ and let $\vec{\theta}$ be the truth-defining formulae. Then, the same translation $\mathcal{T}_{\vec{\theta}}$ is a truth-definition of Q^* in S^*.

Even for Q with extra inference rules, if $\mathcal{T}_{\vec{\theta}}$ further preserves all the inference rules of $\mathsf{Q}[\![\mathsf{B}(P)]\!]$ in the sense that, for each inference rule $\mathcal{R}$ of Q in the form 'if $\phi_0, \dots, \phi_n$ are derived, then infer $\psi_{\vec{\phi}}$', $\mathcal{T}_{\vec{\theta}}$ satisfies that

$$\text{if } \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi_0), \dots, \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\phi_n), \text{ then } \mathsf{S} \vdash \mathcal{T}_{\vec{\theta}}(\psi_{\vec{\phi}}),$$

then Q^* is truth-definable in S^* via $\mathcal{T}_{\vec{\theta}}$.

Proof. By induction on the length of derivation. The $\mathcal{T}_{\vec{\theta}}$-translation of each instance of full induction for $\mathcal{L}_{\mathsf{Q}}(P)$ and each Q^*-axiom trivially holds in S^*. We show that the inference rules are preserved by $\mathcal{T}_{\vec{\theta}}$. Generalization and Modus Ponens are obviously preserved, and so are the other extra rules (if any) by the assumption. For P-Subst, suppose $\mathsf{Q}^* \vdash \phi(P)$. Then, we have $\mathsf{S}^* \vdash \mathcal{T}_{\theta}\big(\phi(P)\big)$ by IH. Since $\mathcal{T}_{\theta}$ does not change P, we have $\mathsf{S}^* \vdash \big(\mathcal{T}_{\vec{\theta}}(\phi)\big)(P)$. Thus, for any $\psi \in \mathcal{L}_{\mathsf{Q}}(P)$, we obtain $\mathsf{S}^* \vdash \big(\mathcal{T}_{\vec{\theta}}(\phi)\big)\big(\widehat{\mathcal{T}_{\vec{\theta}}(\psi)}\big)$ by P-Subst. Finally, since $\mathcal{T}_{\vec{\theta}}$ preserves the logical connectives and quantifiers,

$\mathsf{S}^* \vdash \mathcal{T}_{\vec{\theta}}\big(\phi(\hat{\psi})\big)$ is obtained. □

This property is neither possessed by relative interpretability nor proof-theoretic reducibility. $\mathsf{KF}^*[\![\mathsf{PA}]\!]$ is neither proof-theoretically reducible nor relatively interpretable to $(\mathsf{RT}_{<\varepsilon_0})^*[\![\mathsf{PA}]\!]$ (by Lemma 3.5.6 below), whereas $\mathsf{KF}[\![\mathsf{PA}(P)]\!]$ is both proof-theoretically reducible to and interpretable (by Theorem 3.1.1) in $\mathsf{RT}_{<\varepsilon_0}[\![\mathsf{PA}(P)]\!]$.

Remark 7. If a system B meets the conditions (a)-(d) in §2.1 for our base system, then so does $\mathsf{B}(P)$. Hence, in the setting of the present thesis, whenever we establish the truth definability of Q in S without specifying the base system, it applies to both base systems B and $\mathsf{B}(P)$ and thus Q^* becomes automatically truth-definable in S^* as well by the last lemma. This is the reason why I didn't fix PA as the base system and rather adopt a slightly general setting (cf. Ch.2, §1).

3.3 Digression – Proof-Theoretic Ordinal –

The notion of proof-theoretic ordinal figures prominently among measures of the strength of subsystems of second-order arithmetic and various other systems, and the subject called ordinal analysis is now one of the major subdisciplines of proof system (see [69] or [66] for an overview). However, the usual definition of proof-theoretic ordinal cannot be directly applied to systems of truth. Some alternative definitions still seem implausible for systems of truth. Thus the 'proof-theoretic ordinal' cannot be an appropriate measure until some suitable definition is given for systems of truth. We shortly digress to discuss this issue.

Given a subsystem of second-order arithmetic Q, the proof-theoretic ordinal of Q usually denotes its Π^1_1-ordinal: i.e.,

$$\sup\{\alpha \mid \mathsf{Q} \vdash TI(\prec) \wedge \mathit{otyp}(\prec) = \alpha \wedge \prec \text{ is primitive recursive}\},$$

where $TI(\prec)$ expresses the transfinite induction along $\prec$, that is,

$$\forall X\big[\forall x\big(\forall y \prec x(y \in X) \to x \in X\big) \to \forall x \in field(\prec)\big(x \in X\big)\big].$$

This definition presupposes that the language of Q contains second-order parameters. Since they are not contained in the language of truth systems, the above definition should be modified in a certain way.

One such modification for systems without second-order parameters can be found in [69] for example. There Rathjen considers augmenting a system with a new unary predicate U and redefine

$$TI(\prec) := \forall x\big(\forall y \prec xU(y) \to U(x)\big) \to \forall x \in field(\prec)U(x)\big].$$

However, it remains to be decided which properties are to be assumed for U, and indeed the resulting ordinal depends on which properties are assumed. Suppose that we postulate all the axioms governing the truth predicate and mathematical induction even for U. Then, for instance, the resulting proof-theoretic ordinal of $\mathsf{KF}[\![\mathsf{PA}]\!]$ would be $\varphi_{\varepsilon_0}(0)$. On the other hand, when we only postulate induction for U, the proof theoretic ordinal would be ε_0.[7] If we postulate nothing for U then the ordinal would then be ω.

Another possible definition is for it to be the α such that $\mathsf{Q} \equiv \mathrm{PA} + \bigcup_{\beta<\alpha}\mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(\beta)$ or $\sup\{\beta \mid$

[7] I only give a sketch of its proof here. Let us denote the system we are considering by $\mathsf{KF}(U)$. Take a suitable semi-formal system to which $\mathsf{KF}(U)$ is embedded; such a system can be obtained by adding inference rules for the truth predicate T to ω-arithmetic; cf. [5]. In this system, no sentence containing U can be critical in the rules for T. We call a derivation $\mathcal{D}$ in this system *quasi-normal*, when cut is only applied to atomics of the forms Tt and $\neg Tt$. Then, we can show that, if $\prec$ is a well-ordering and $otyp(\prec)$ is limit, $otyp(\prec)$ is always less than or equal to the least length α of quasi-normal derivations of $TI(\prec)$. This is shown in a similar way to Theorem 6.7.2 (Boundedness Theorem) of [66]. It is crucial in its proof that U cannot be critical in any rule for T. For example, in showing the analogue to Lemma 6.6.8 (Boundedness Lemma) of [66] (we can take a fixed interpretation of T in $\mathbb{N}$, say, the least Kripkean fixed-point, since it doesn't depend on the interpretation of U at all), though T may occur in Δ (we are using the same notation of [66]), the positiveness or negativeness of U's occurrences isn't affected by any rule for T and the induction steps work in the same way as the cited lemma; by contrast, in the case of $\mathsf{KF}[\![\mathsf{PA}(U)]\!]$, this argument is no longer valid because U can be critical in the rules for T and the interpretation of T must depend on the interpretation of U. However, we can also show by the standard technique that each derivation $\mathcal{D}$ in $\mathsf{KF}(U)$ can be transformed into a quasi-normal derivation $\mathcal{D}'$ in the semi-formal system of the length $< \varepsilon_0$. This entails that if $\mathsf{KF}(U) \vdash TI(\prec)$ then $otyp(\prec) < \varepsilon_0$.

$\mathrm{PA} + \mathrm{TI}_{\mathcal{L}_{\mathsf{PA}}}(\beta) \subset \mathsf{Q}\}$. However, both are only focusing on the truth-free part, and thus they are unsatisfactory for the same reason as the reason why we reject proof-theoretic reducibility or conservativity.[8]

3.4 Several Results on Truth Definability

In this section, we will present several results on truth-definability among systems we have introduced in §2.

For simplicity and readability, let $\preceq$ stand for the truth-definability relation: i.e., for systems Q_0 and Q_1, we write $\mathsf{Q}_0 \preceq \mathsf{Q}_1$ if Q_0 is truth-definable in Q_1. Then, $\mathsf{Q}_0 \prec \mathsf{Q}_1$ is for $\mathsf{Q}_0 \preceq \mathsf{Q}_1$ & $\mathsf{Q}_1 \not\preceq \mathsf{Q}_0$.

Some results on truth-definability immediately follow from already known or shown facts. First, if Q_0 is a subsystem of Q_1, then the former is trivially truth-definable in the latter (via the identity translation). Hence, for example,

$$\mathsf{TB} \preceq \mathsf{UTB} \qquad \mathsf{PUTB} \preceq \mathsf{KF} \qquad \mathsf{KF} \preceq \mathsf{KF}^*.$$

Second, since truth-definability implies conservativity for $\mathcal{L}_0$-sentences, any system not conservative (for $\mathcal{L}_0$-sentences) over a system Q is not truth-definable in Q. Hence, for example, the following negative results on truth-definability are already available:

$$\mathsf{TB}, \mathsf{UTB}, \mathsf{TC}^-, \mathsf{KF}^- \not\succeq \mathsf{TC} \not\succeq \mathsf{FS} \not\succeq \mathsf{RT}_{<\varepsilon_0}$$

Third, the next proposition follows from Proposition 3.2.4, since no system but FS is ω-inconsistent.[9]

[8] For another example, Beklemishev [2] introduced Π^0_n-ordinal of systems ($n \in \mathbb{N}$) which can be defined without appealing to second-order parameters. However, this definition should be rejected for the same reason that it only looks at the arithmetical (truth-free) part.

[9] Recall that we have assumed that B is arithmetically sound and thus has a model over $\mathbb{N}$. Then, it is known

Lemma 3.4.1. None of the systems introduced so far (except FS itself) can define the truth of FS.

Fourth, due to Tarski's undefinability theorem, no system of truth presented so far is truth-definable in PA, since every system can derive the TB-schema.

In the rest of this section, we will show less trivial results.

Lemma 3.4.2. $\mathsf{KF}^- \not\preceq \mathsf{TC}^-$, and $\mathsf{TC}^- \not\preceq \mathsf{KF}^-$. On the other hand, we have $\mathsf{TC} \prec \mathsf{KF}$ since the former is a subsystem of the latter.

Proof. The first claim follows from the fact that KF is stronger than TC (cf. Theorem 3.4.4 below). For the second claim, let B' be B plus full-induction for $\mathcal{L}_0$. For the sake of contradiction, suppose $\mathsf{TC}^- \preceq \mathsf{KF}^-$. Since a relative truth definition preserves the $\mathcal{L}_0$-part, this entails that $\mathsf{TC}^-[\![\mathsf{B}']\!] \preceq \mathsf{KF}^-[\![\mathsf{B}']\!]$ as well. However, every nonstandard model $M \models \mathsf{B}'$ must be recursively saturated with respect to $\mathcal{L}_0$ (i.e., every recursive type $p(x)$ from $\mathcal{L}_0$ over M is realized in M) whenever it is expandable to a model of $\mathsf{TC}^-[\![\mathsf{B}']\!]$; Lachlan [51] showed this for $\mathsf{B}' = \mathsf{PA}$, but the proof of Lachlan's result (also see [47]) can be directly generalized to an arbitrary base system B' over $\mathcal{L}_0$ which meets the conditions we posed in §2 and contains full-induction for $\mathcal{L}_0$. On the other hand, since B is arithmetically sound and thus B' has an infinite model, B' has a non-recursively saturated model; we can take the prime model of some suitable complete extension of B', which is not recursively saturated but a nonstandard model of this extension.[10] By Proposition 3.2.4 (1), these contradict the fact that any model $M \models \mathsf{B}'$ can be expanded to a model of $\mathsf{KF}^-[\![\mathsf{B}']\!]$ ([5, §5]). □

Lemma 3.4.3. $\mathsf{TB}, \mathsf{UTB} \prec \mathsf{TC}^-$

Proof. It is obvious that $\mathsf{TB}, \mathsf{UTB} \preceq \mathsf{TC}^-$, since the former two are subsystems of the latter. We can show by the same argument as the last lemma that the converse fails, since TB and UTB are

that all the systems in §2 except FS has a model over $\mathbb{N}$ and thus ω-consistent.

[10] For the details of the proof, we refer the readers to [47, §8.1, §11.2], where these results are proved for PA but they can be generalized for an arbitrary B with full-induction.

conservative over B in the semantic sense.[11] $\square$

There still remains one open problem on the relative truth definability among conservative systems of truth over B: we do not yet know whether UTB is truth-definable in TB.

Theorem 3.4.4. $\mathsf{RT}_{<\varepsilon_0}$ is truth-definable in KF.

We prove a more general fact from which this theorem easily follows. Preliminarily, define an $\mathcal{L}_T$-formula $D^+(x)$ by

$$D^+(x) := \mathrm{St}_{\mathcal{L}_T}(x) \wedge (Tx \vee Fx) \wedge \neg(Tx \wedge Fx).$$

Notice that $D^+(x)$ is equivalent to $(Fx \leftrightarrow \neg Tx) \wedge x \in \mathrm{St}_{\mathcal{L}_T}$.

Lemma 3.4.5. Let Q be a system over $\mathcal{L}_T$ which derives the following:

(i) The axiom **K1** (of KF);

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

(iii) $[\mathrm{St}_{\mathcal{L}_T}(\dot{\forall} z.x) \wedge \forall y D^+ x(y/z)] \rightarrow D^+(\dot{\forall} z.x)$;

(iv) $D^+(x) \rightarrow (T\neg x \leftrightarrow \neg Tx)$[12];

(v) $D^+(x \dot{\vee} y) \rightarrow [Tx \dot{\vee} y \leftrightarrow (Tx \vee Ty)]$;

(vi) $D^+(x \dot{\rightarrow} y) \rightarrow [Tx \dot{\rightarrow} y \leftrightarrow (Tx \rightarrow Ty)]$;

(vii) $D^+(\dot{\forall} z.x) \rightarrow [T\dot{\forall} z.x \leftrightarrow \forall y Tx(y/z)]$;

(viii) The axiom **K2** (of KF).

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

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

[11] Let B' be B plus full-induction for $\mathcal{L}_0$. We can show that UTB plus full-induction for $\mathcal{L}_T$ is proof-theoretically reducible to B' for $\mathcal{L}_0$ by using partial truth predicates; cf. fn.5. This gives an alternative proof of this lemma, since $\mathsf{TC}^-[\![\mathsf{B}]\!] \preceq \mathsf{UTB}[\![\mathsf{B}]\!]$ implies $\mathsf{TC}^-[\![\mathsf{B}']\!] \preceq \mathsf{UTB}[\![\mathsf{B}']\!]$ and thus $\mathsf{TC}[\![\mathsf{B}']\!]$ would become truth-definable in $\mathsf{UTB}[\![\mathsf{B}']\!]$ plus full-induction for $\mathcal{L}_T$. In contrast, it is known that UTB plus full induction is not conservative over B' in the semantic sense; only recursively saturated B'-models can be expanded to its model (for the proof, see [47, §15]).

[12] This (iv) is actually redundant since it follows from the definition of D^+ as was noted.

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

We write $h_\beta(x)$ for $h(x,\beta)$. Then, by the primitive recursion theorem, we take another primitive recursive function k such that:

$$k(a) := \begin{cases} a & \text{if } a \in \mathrm{AtFml}_{\mathcal{L}_0} \\ \ulcorner T\big(k \circ h_\gamma(b)\big) \urcorner & \text{if } a = \dot{T}_\gamma b \text{ for } b \in \mathrm{Term} \\ \dot{\neg} \ulcorner Tk\dot{b} \urcorner & \text{if } a = \dot{\neg} b \\ \ulcorner Tk\dot{b} \urcorner \dot{\vee} \ulcorner Tk\dot{c} \urcorner \ \big(\ulcorner Tk\dot{b} \urcorner \dot{\rightarrow} \ulcorner Tk\dot{c} \urcorner, \text{ resp.}\big) & \text{if } a = b \dot{\vee} c \text{ (or } a = b \dot{\rightarrow} c) \\ \dot{\forall} c. \ulcorner Tk(\dot{b}(c/\ulcorner c \urcorner)) \urcorner & \text{if } a = \dot{\forall} c.b \text{ for } c \in \mathrm{Var} \\ \ulcorner 0 = 1 \urcorner & \text{otherwise;} \end{cases}$$

We slightly abuse notation here; $\ulcorner T\big(k \circ h_\gamma(b)\big) \urcorner$ and $\dot{\forall} z. \ulcorner Tk(\dot{b}(c/\ulcorner c \urcorner)) \urcorner$ precisely mean $\ulcorner T\big(k \circ h_\gamma(u)\big) \urcorner (b/\ulcorner u \urcorner)$ and $\dot{\forall} c. \ulcorner Tk(\dot{b}(u/\ulcorner c \urcorner)) \urcorner (c/\ulcorner u \urcorner)$ where u is a fresh variable; informally, $k(T_\gamma t) = T\big(k \circ h_\gamma(t)\big)$ and $k(\forall z \phi(z)) = \forall z T\big(k \ulcorner \phi(\dot{z}) \urcorner\big)$ for a term t and formula ϕ. We notice that if $x \in \mathrm{St}_\beta$ for $\beta < \alpha$ then $kx \in \mathrm{St}_{\mathcal{L}_T}$. Then, for each $\beta < \alpha$, we set $\theta_\beta(x)$ to be $Tk(x)$ and we take $\theta_\beta(x)$ as the truth-defining formula of T_β.

We will first show by $\mathrm{TI}_{\mathcal{L}_T}(\beta)$ for each fixed $\beta < \alpha$ that

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

Since $\mathsf{Q} \vdash \mathrm{TI}_{\mathcal{L}_T}(< \alpha)$, it suffices to show that $\mathrm{St}_\gamma(x) \to D^+(kx)$ is progressive on γ. Let $a \in \mathrm{St}_\gamma$ and suppose the claim up to γ. We show the claim for γ by subinduction on the surface complexity of the sentence coded by a.

The case in which $a \in \mathrm{AtSent}_{\mathcal{L}_0}$ immediately follows from (i) since $k(a) = a$. Suppose $a = \dot{T}_\delta x$

for $x \in \mathrm{CT}$ and $\delta < \gamma$. Then, by (viii), it is observed that:

$$D^+(ka) \Leftrightarrow \big[T^\ulcorner Tk \circ h_\delta(x)^\urcorner \vee F^\ulcorner Tk \circ h_\delta(x)^\urcorner\big] \wedge \big[\neg T^\ulcorner Tk \circ h_\delta(x)^\urcorner \vee \neg F^\ulcorner Tk \circ h_\delta(x)^\urcorner\big]$$
$$\Leftrightarrow \big[Tk \circ h_\delta(\mathrm{val}(x)) \vee Fk \circ h_\delta(\mathrm{val}(x))\big] \wedge \big[\neg Tk \circ h_\delta(\mathrm{val}(x)) \vee \neg Fk \circ h_\delta(\mathrm{val}(x))\big].$$

If $\mathrm{val}(x) \in \mathrm{St}_\delta$, then $h_\delta(\mathrm{val}(x)) = \mathrm{val}(x)$ and thus $k \circ h_\delta(\mathrm{val}(x)) = k(\mathrm{val}(x))$; the claim follows from IH. Otherwise, $k \circ h_\delta(\mathrm{val}(x)) = {}^\ulcorner 0{=}1^\urcorner$; the claim trivially follows from (i). For the cases in which $a = \dot\neg b$, $a = b \dot\vee c$ or $a = b \dot\to c$, we first observe that $D^+(kx) \leftrightarrow D^+({}^\ulcorner Tk\dot{x}^\urcorner)$ for all $x \in \mathrm{St}_\gamma$, which is shown by (viii). Then the claim follows from (ii). Next, suppose a is of the form $\dot\forall c.b$. We already have $(\forall y) D^+\big(k(b(y/c))\big)$ by SIH, and we also have:

$$D^+\big({}^\ulcorner Tk(\dot{b}(c/{}^\ulcorner c^\urcorner))^\urcorner(y/c)\big) \Leftrightarrow D^+\big({}^\ulcorner Tk(\dot{b}(\dot{y}/{}^\ulcorner c^\urcorner))^\urcorner\big) \Leftrightarrow D^+k(b(y/c))$$

for each y in Q. This entails the claim $D^+(\dot\forall c.{}^\ulcorner Tk(\dot{b}(c/{}^\ulcorner c^\urcorner))^\urcorner)$ by (iii). Thus, we have shown (3.2).

Now we can straightforwardly show using (iv)-(viii) that $\mathcal{T}_{\vec\theta}$ is indeed a truth definition of $\mathsf{RT}_{<\alpha}$ in Q. We illustrate $\mathsf{Q} \vdash \mathcal{T}_{\vec\theta}(\mathbf{R2})$ and $\mathsf{Q} \vdash \mathcal{T}_{\vec\theta}(\mathbf{R6})$ for examples. For the former, let $\beta < \alpha$ and $x \in \mathrm{St}_\beta$. Then $\mathcal{T}_{\vec\theta}(T_\beta \dot\neg x)$ is equivalent to $T(\dot\neg {}^\ulcorner Tk\dot{x}^\urcorner)$. By (3.2), we already know both $D^+(\dot\neg {}^\ulcorner Tk\dot{x}^\urcorner)$ and $D^+(kx)$. Hence,

$$T(\dot\neg {}^\ulcorner Tk\dot{x}^\urcorner) \Leftrightarrow \neg T({}^\ulcorner Tk\dot{x}^\urcorner) \Leftrightarrow \neg Tkx = \mathcal{T}_{\vec\theta}(\neg T_\beta x);$$

we thus obtain $\mathsf{Q} \vdash \mathcal{T}_{\vec\theta}(\mathbf{R2})$. Next, for the latter, let $\beta < \alpha$ and take any $\delta < \beta$ and $x \in \mathrm{CT}$ with $\mathrm{val}(x) \in \mathrm{St}_\delta$. Then, it follows from (viii) that

$$\mathcal{T}_{\vec\theta}\big((T_\beta(\dot{T}_\delta x)\big) \;\Leftrightarrow\; T\big({}^\ulcorner Tk \circ h_\delta x^\urcorner\big) \;\Leftrightarrow\; Tk \circ h_\delta(\mathrm{val}(x))$$

Since we have assumed that $\mathrm{val}(x) \in \mathrm{St}_\delta$, we have $k \circ h_\delta(\mathrm{val}(x)) = k(x)$ and thus the last formula is equivalent to $\mathcal{T}_{\vec{\theta}}\big(T_\beta \mathrm{val}(x)\big) = Tk(\mathrm{val}(x))$. The other cases can be shown similarly. □

Since KF^- satisfies the conditions (i)-(viii)[13] and $\mathsf{KF} \vdash \mathrm{TI}_{\mathcal{L}_T}(< \varepsilon_0)$ (by the standard proof), Theorem 3.4.4 follows from this lemma. In addition, FKF^- also meets all the conditions (i)-(viii), the next theorem follows.

Theorem 3.4.6. $\mathsf{RT}_{<\varepsilon_0}$ is truth-definable in FKF (and thus DT).

Corollary 3.4.7. If $\alpha \leq \varepsilon_0$, then $\mathsf{DT}^-, \mathsf{KF}^- \not\preceq \mathsf{RT}_{<\alpha}$.[14]

Proof. We only show the claim for KF^-, but DT^- case is shown in the same manner. When $\alpha < \varepsilon_0$, the claim is obvious since $\mathsf{RT}_{<\varepsilon_0} \leq \mathsf{KF}$. Let $\alpha = \varepsilon_0$. For the sake of contradiction, suppose KF is truth-definable in $\mathsf{RT}_{<\varepsilon_0}$. Let $\theta(x)$ be the $\mathcal{L}_{\varepsilon_0}$-formula defining the truth of KF. Take a finite $\Gamma \subset \mathsf{RT}_{<\varepsilon_0}$ such that $\Gamma \vdash \mathcal{T}_\theta(\mathbf{K0}) \wedge \cdots \wedge \mathcal{T}_\theta(\mathbf{K8})$. Take the maximum $\beta < \varepsilon_0$ such that T_β occurs in Γ or θ. Then, we have $\Gamma \cup \{\theta\} \subset \mathsf{RT}_{<\beta+1}$ and thus KF would be truth-definable in $\mathsf{RT}_{<\beta+1}$; however, since $\beta + 1 < \varepsilon_0$, it leads to a contradiction. □

These techniques can be modified for schematic reflective closures.

Lemma 3.4.8. Suppose Q over $\mathcal{L}_T$ satisfies (i)-(viii) of Lemma 3.4.5. Then, $\mathsf{RT}_{<\Gamma_0}$ is truth-definable in Q^*. We can take the same truth-definition $\mathcal{T}_\theta$ as in the proof of Lemma 3.4.5.

Proof. Define ordinals β_n for each $n \in \mathbb{N}$ by $\beta_0 = \varepsilon_0$ and $\beta_{n+1} = \varphi_{\beta_n}(0)$. Note that $\Gamma_0 = \lim_{n\to\infty} \beta_n$ and β_n is an epsilon number for all $n \in \mathbb{N}$. By Lemma 3.4.5, it suffices to show that $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_T(P)}(< \beta_n)$ for all $n \in \mathbb{N}$. This is shown by induction on n.

Recall that Q^* contains full induction for $\mathcal{L}_T(P)$. Thus $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_T(P)}(< \varepsilon_0)$: we have shown the base step. For the induction step, suppose $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_T(P)}(< \beta_n)$. Then, by Lemma 3.4.5,

[13] See Lemma 3.2.4 and 3.2.5 of [17], where $D^+(x)$ is denoted by $D(x)$.

[14] We might expect that ε_0 can be replaced by Γ_0 in this corollary. I do not yet know whether it is true or false, but we can at least show that there is no truth definition of $\mathsf{KF}^-[\![\mathsf{B}(P)]\!]$ (and $\mathsf{DT}^-[\![\mathsf{B}(P)]\!]$) in $\mathsf{RT}_{<\Gamma_0}[\![\mathsf{B}(P)]\!]$. For there would otherwise be a truth definition KF^* in $(\mathsf{RT}_{<\Gamma_0})^*$ and then we could derive a contradiction in a similar way to the below proof by using Lemmata 3.4.8 and 3.5.6.

$\mathsf{RT}_{<\beta_n}[\![\mathsf{B}(P)]\!]$ is truth-definable in Q^*. It follows from Theorem 2.3.8 that $\mathsf{RT}_{<\beta_n}[\![\mathsf{B}(P)]\!] \vdash \mathrm{TI}_{\mathcal{L}_0(P)}(< \beta_{n+1})$ and thus $\mathsf{Q}^* \vdash \mathrm{TI}_{\mathcal{L}_0(P)}(< \beta_{n+1})$. In particular $\mathsf{Q}^* \vdash \mathrm{TI}\big(< \beta_{n+1}; P(x)\big)$. Finally, we apply P-Subst and obtain $\mathrm{TI}_{\mathcal{L}_T(P)}(< \beta_{n+1})$. □

Theorem 3.4.9. $\mathsf{RT}_{<\Gamma_0}$ is truth-definable in KF^* and FKF^* (and thus in DT^*).

We now turn to the relative truth definability among type-free systems of truth. Cantini [5] showed that KF can construct a fixed-point of a positive arithmetical operator. Looking closely into his proof, it is observed that only the PUTB schema is needed for this construction. Let $\mathcal{L}_0^{(2)}$ be the second-order language which expands $\mathcal{L}_0$ with second-order variables. We have the following.

Lemma 3.4.10 (Cantini [5])**.** Let $\Phi(X, x, \vec{v})$ be a $\mathcal{L}_0^{(2)}$-formula with no second-order quantifier and with only displayed variables free in which X occurs only positively. Then, there exists a $\mathcal{L}_T$-formula $\phi(x, \vec{v})$ such that

$$\mathsf{PUTB}^- \vdash \phi(x, \vec{v}) \leftrightarrow \Phi(\hat{u}\phi(u, \vec{v}), x, \vec{v}),$$

where $\Phi(\hat{u}\phi(u, \vec{v}), x, \vec{v})$ is the result of substituting $\phi(t, \vec{v})$ for each occurrence of $t \in X$ in Φ.

Proof. By parametrized diagonalization (applied to a $\mathcal{L}_T$-formula $\Phi(\hat{u}Ta(u/\ulcorner x \urcorner, \vec{v}/\ulcorner \vec{v} \urcorner), x, \vec{v})$ w.r.t. the variable a), we can find a γ such that

$$\mathsf{PUTB}^- \vdash \gamma(x, \vec{v}) \leftrightarrow \Phi(\hat{u}T\ulcorner\gamma(\dot{u}, \dot{\vec{v}})\urcorner, x, \vec{v})$$

From the construction of the diagonal formula γ, we can assume that γ is T-positive. Hence, the PUTB schema implies

$$\mathsf{PUTB}^- \vdash \gamma(x, \vec{v}) \leftrightarrow \Phi(\hat{u}\gamma(u, \vec{v}), x, \vec{v})$$

We can take this γ as the desired ϕ. $\square$

It is observed by Feferman [17] that the truth predicate of KF can be defined by a fixed-point of a certain positive arithmetical operator; Feferman made use of this fact to show that Σ^1_1-AC, within which a fixed-point (but not necessarily the least) of any positive Σ^1_1-operator can be constructed, can define the truth of KF. Since full induction isn't used in Feferman's construction, we indeed carry out the same argument in PUTB^-:

Theorem 3.4.11. KF^- is truth-definable in PUTB^-.

Recall that, since a base system B is arbitrary and thus we can replace B by $\mathsf{B}(P)$, truth-definability carries over to schematic reflective closures (also see the Remark in §3.3).

Corollary 3.4.12. KF and KF^* are truth-definable in PUTB and PUTB^* respectively.

Theorem 3.4.13. $\mathsf{PUTB}^- \equiv \mathsf{B}$, $\mathsf{PUTB} \equiv \mathsf{KF}$ and $\mathsf{PUTB}^* \equiv \mathsf{KF}^*$. In particular, we have $\mathsf{PUTB}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$.

It follows from Feferman's consistency proof of DT in [19] that the truth of FKF^- is also defined as an arbitrary fixed-point of a certain positive arithmetical operator. Thus, FKF and FKF^- are truth definable in PUTB and PUTB^- respectively. However, as we have noted in §2.5, fixed-points constructible in $\mathsf{PUTB}^{(-)}$ (or Σ^1_1-AC) are not necessarily consistent. Thus, these techniques are not directly applied to obtain that systems with Cons are truth-definable in $\mathsf{PUTB}^{(-)}$. Fortunately, we can nonetheless modify these techniques so that PUTB^- can define the truth of $\mathsf{FKF}^- + \mathrm{Cons}$, i.e., DT^-.[15]

Theorem 3.4.14. DT^- is truth-definable in KF^- (and thus in PUTB^-).

Proof. First, using the recursion theorem, we define (in PA) a preliminary translation I (on the codes of formulae) such that:

[15] In contrast, it is still open whether PUTB^- can define the truth of $\mathsf{KF}^- + \mathrm{Cons}$.

$$I(\phi) = \begin{cases} \phi & \text{if } \phi \text{ is } t = s \text{ or } t \neq s, \\ T\dot{I}t & \text{if } \phi \text{ is } Tt, \\ T\dot{I}\dot{\neg}t & \text{if } \phi \text{ is } \neg Tt, \\ I(\psi) & \text{if } \phi \text{ is } \neg\neg\psi \\ [I\psi_0 \wedge I\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\neg\psi_1] & \\ \quad \vee [I\neg\psi_0 \wedge I\psi_1 \wedge \neg I\psi_0 \wedge \neg I\neg\psi_1] & \\ \quad \vee [I\psi_0 \wedge I\neg\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\psi_1] & \text{if } \phi \text{ is } \psi_0 \vee \psi_1 \\ I\neg\psi_0 \wedge I\neg\psi_1 \wedge \neg I\psi_0 \wedge \neg I\psi_1 & \text{if } \phi \text{ is } \neg(\psi_0 \vee \psi_1) \\ [I\psi_0 \wedge I\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\neg\psi_1] \vee [I\neg\psi_0 \wedge \neg I\psi_0] & \text{if } \phi \text{ is } \psi_0 \to \psi_1, \\ I\psi_0 \wedge I\neg\psi_1 \wedge \neg I\neg\psi_0 \wedge \neg I\psi_1 & \text{if } \phi \text{ is } \neg(\psi_0 \to \psi_1), \\ \forall x(I\psi \wedge \neg I\neg\psi) & \text{if } \phi \text{ is } \forall x\psi, \\ \forall x\big((I\psi \wedge \neg I\neg\psi) \vee (I\neg\psi \wedge \neg I\psi)\big) \wedge \exists x I\neg\psi & \text{if } \phi \text{ is } \neg\forall x\psi \\ 0 & \text{otherwise,} \end{cases}$$

where $\dot{I}$ is the representation of I. Let $\theta(x)$ be

$$TIx \wedge \neg TI\dot{\neg}x \wedge FI\dot{\neg}x \wedge \neg FIx.$$

Then we will show that this θ defines the truth of $\mathsf{FKF}^- + \text{Cons}$ in KF^-. This can be done in a straightforward but somewhat tedious manner; we will show some typical cases.

First, note that $\mathcal{T}_\theta(Ft)$ is $TI\dot{\neg}t \wedge \neg TIt \wedge FIt \wedge \neg FI\dot{\neg}t$. Then, it immediately follows that $\mathsf{KF}^- \vdash \mathcal{T}_\theta(\text{Cons}) \wedge \mathcal{T}_\theta(\mathbf{K2})$. By definition of I, we obviously have $\mathsf{KF}^- \vdash \mathcal{T}_\theta(\mathbf{K1}) \wedge \mathcal{T}_\theta(\mathbf{K3})$ as well.

For **FK4**, let $x, y \in \mathrm{St}_{\mathcal{L}_T}$. Then, $\mathcal{T}_\theta(Tx \veebar y)$ is equivalent to

$$\begin{aligned}
&\big[(TIx \wedge TIy \wedge FI\dot{\neg}x \wedge FI\dot{\neg}y) \vee (TI\dot{\neg}x \wedge TIy \wedge FIx \wedge FI\dot{\neg}y) \\
&\qquad \vee (TIx \wedge TI\dot{\neg}y \wedge FI\dot{\neg}x \wedge FIy)\big] \\
&\wedge \big[\neg TI\dot{\neg}x \vee \neg TI\dot{\neg}y \vee \neg FIx \vee \neg FIy\big] \\
&\wedge \big[FI\dot{\neg}x \vee FI\dot{\neg}y \vee TIx \vee TIy\big] \\
&\wedge \big[(\neg FIx \wedge \neg FIy \wedge \neg TI\dot{\neg}x \wedge \neg TI\dot{\neg}y) \vee (\neg FI\dot{\neg}x \wedge \neg FIy \wedge \neg TIx \wedge \neg TI\dot{\neg}y) \\
&\qquad \vee (\neg FIx \wedge \neg FI\dot{\neg}y \wedge \neg TI\dot{\neg}x \wedge \neg TIy)\big]
\end{aligned}$$

Since the second and third clauses are entailed from the fourth and first respectively, they can be dropped. Then, the resulting formula is equivalent to:

$$\begin{aligned}
&\big[TIx \wedge TIy \wedge FI\dot{\neg}x \wedge FI\dot{\neg}y \wedge \neg FIx \wedge \neg FIy \wedge \neg TI\dot{\neg}x \wedge \neg TI\dot{\neg}y\big] \\
&\quad \vee \big[TI\dot{\neg}x \wedge TIy \wedge FIx \wedge FI\dot{\neg}y \wedge \neg FI\dot{\neg}x \wedge \neg FIy \wedge \neg TIx \wedge \neg TI\dot{\neg}y\big] \\
&\quad \vee \big[TIx \wedge TI\dot{\neg}y \wedge FI\dot{\neg}x \wedge FIy \wedge \neg FIx \wedge \neg FI\dot{\neg}y \wedge \neg TI\dot{\neg}x \wedge \neg TIy)\big].
\end{aligned}$$

This formula is equivalent to $\mathcal{T}_\theta\big((Tx \wedge Ty) \vee (Tx \wedge Fy) \vee (Fx \wedge Ty)\big)$. We finally obtain $\mathsf{KF}^- \vdash \mathcal{T}_\theta(\mathbf{FK4})$. □

Theorem 3.4.15. DT and DT^* are truth-definable in KF and KF^* respectively (and thus in PUTB and PUTB^* respectively).

From the truth-definability of PUTB^- and DT^- in KF^-, we know:

Corollary 3.4.16. If B is reflexive, then PUTB^- and DT^- are reflexive. PUTB and DT are also (essentially) reflexive since they contain full induction.

However, it still remains open whether the converses of Theorem 3.4.15 hold. Concluding this

section, we restate the two open problems for future work.

Open Problem 1: Is PUTB^- (or PUTB) truth-definable in DT^- (DT, resp.)?

Open Problem 2: Is UTB truth-definable in TB?

3.5 Some Applications

In the present section, we exhibit some applications of the results and techniques on relative truth definability developed so far. First of all, as we have announced, we give positive answers to both the first and the second conjectures of Feferman.

Theorem 3.5.1. $\mathsf{DT}[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\varepsilon_0}$ and $\mathsf{DT}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\Gamma_0}$.

Proof. The upper bound of $\mathsf{DT}[\![\mathsf{PA}]\!]$ immediately follows from Theorem 3.4.15. The converse follows from Theorems 2.3.7 and 3.4.6.

Similarly, the upper bound of $\mathsf{DT}^*[\![\mathsf{PA}]\!]$ is also obtained by Theorem 3.4.15, and it follows from Theorem 2.3.7 and 3.4.9 that this bound is exact. □

As the next application, we will show some proof-theoretic equivalences among systems of iterative compositional truth.

Feferman [17, p.28] suggested that the proof-theoretic strength of $\mathsf{WKF}[\![\mathsf{PA}]\!]$ is at least as strong as $\mathsf{RA}_{<\varepsilon_0}$. This is indeed correct since WKF meets the conditions (i)-(viii) of Lemma 3.4.5 and proves $\mathrm{TI}_{\mathcal{L}_T}(< \varepsilon_0)$. For the converse, we can show that a simple interpretation which translates '$T(x \dot{\rightarrow} y)$' by '$T(\dot{\neg} x \dot{\vee} y)$' is a truth-definition of WKF^- (+Cons) in FKF^- (+Cons, resp.); note that we use the recursion theorem to define this translation. Hence, by Theorems 2.4.2 and 3.4.15, we know the bound is exact. Consequently, we have the following:

Theorem 3.5.2. (1) WKF^- and $\mathsf{WKF}^- + \mathrm{Cons}$ are truth-definable in KF^-.

(2) $\mathsf{WKF}^- \equiv \mathsf{WKF}^- + \mathrm{Cons} \equiv \mathsf{B}$.

(3) $\mathsf{WKF}[\![\mathsf{PA}]\!] \equiv \mathsf{WKF}[\![\mathsf{PA}]\!] + \mathrm{Cons} \equiv \mathsf{RA}_{<\varepsilon_0}$.

(4) $\mathsf{WKF}^*[\![\mathsf{PA}]\!] \equiv \mathsf{WKF}^*[\![\mathsf{PA}]\!] + \mathrm{Cons} \equiv \mathsf{RA}_{<\Gamma_0}$.

We have so far focused on Cons. Then what would happen when we add Comp? As a matter of fact, Comp collapses all the differences among KF, FKF and WKF: that is, we have the following.

Lemma 3.5.3. $\mathsf{KF}^- + \mathrm{Comp}$, $\mathsf{FKF}^- + \mathrm{Comp}$ and $\mathsf{WKF}^- + \mathrm{Comp}$ are all identical.

Proof. The proof is straightforward. For example, observe that $Tx \vee Ty$ is equivalent to $(Tx \wedge Fx) \vee (Fx \wedge Ty) \vee (Tx \wedge Ty)$ under Comp. □

In the rest of the present section, we will determine the proof-theoretic strength of the schematic reflective closures of $\mathsf{RT}_{<\alpha}$ and FS by using the results and techniques on truth-definability developed so far. As we shall see, schematic reflective closure does not increase proof-theoretic strength in the cases of FS and $\mathsf{RT}_{<\alpha}$ with $\alpha = \omega^\xi$ for some $\xi > 0$.

For each $k \in \mathbb{N}$ and ordinal α, $(\mathsf{RT}_{<\alpha})^* \restriction_k$ denotes the system obtained from $(\mathsf{RT}_{<\alpha})^*$ by restricting the number of applications of P-Subst in each derivation to at most k times; therefore $(\mathsf{RT}_{<\alpha})^* \restriction_0$ is just identical with $\mathsf{RT}_{<\alpha}[\![\mathsf{B}(P)]\!]$. In the following, we assume that derivations are formulated in Hilbert style in which inference rules are Modus Ponens and P-Subst only.

Lemma 3.5.4. $(\mathsf{RT}_{<\alpha})^* \restriction_{k+1}$ is proof-theoretically reducible to $(\mathsf{RT}_{<\alpha+\alpha})^* \restriction_k$ for $\mathcal{L}_\alpha(P)$.

Proof. Take an arbitrary derivation $\mathcal{D} = \langle \theta_0, \ldots, \theta_n \rangle$ in $(\mathsf{RT}_{<\alpha})^* \restriction_{k+1}$. Let $\mathcal{D}_0 = \langle \theta_0, \ldots, \theta_m \rangle$ $(m \leq n)$ be the initial segment of $\mathcal{D}$ which ends with the first application of P-Subst $\frac{\theta_j(P)}{\theta_j(\hat{\psi})}$ for some $\psi \in \mathcal{L}_\alpha(P)$: thus θ_m is $\theta_j(\hat{\psi})$ $(j < m)$. Then, there exist $\beta, \gamma \leq \alpha$ such that $\theta_0, \ldots, \theta_m \in \mathcal{L}_\beta$ and $\psi \in \mathcal{L}_\gamma$. We define a translation $\mathcal{S}^\beta_\psi$ from $\mathcal{L}_\beta(P)$ to $\mathcal{L}_{\gamma+\beta}(P)$, which 'lifts up' the level of truth predicates in each θ_i $(i < m)$ so that ψ can be treated as if it were at the lowest level; note that

$\gamma + \beta \leq \alpha + \alpha$. First we define a preliminary translation g^{β}_{ψ} (on the codes of formulae) by the recursion theorem such that:

$$g^{\beta}_{\psi}(\phi) := \begin{cases} \phi & \text{if } \phi \text{ is an } \mathcal{L}_0\text{-atomic} \\ \psi(t) & \text{if } \phi \text{ is } P(t) \\ T_{\gamma+\delta}\big(\dot{g}^{\beta}_{\psi}(t)\big) & \text{if } \phi \text{ is } T_{\delta}(t) \\ \neg g^{\beta}_{\psi}(\theta) & \text{if } \phi \text{ is } \neg\theta \\ g^{\beta}_{\psi}(\theta_0) \vee g^{\beta}_{\psi}(\theta_1) \ \big(g^{\beta}_{\psi}(\theta_0) \to g^{\beta}_{\psi}(\theta_1), \text{resp.}\big) & \text{if } \phi \text{ is } \theta_0 \vee \theta_1 \text{ (or } \theta_0 \to \theta_1) \\ \forall z. g^{\beta}_{\psi}(\theta) & \text{if } \phi \text{ is } \forall z \theta(z), \end{cases}$$

where $\dot{g}^{\beta}_{\psi}$ is the representation of g^{β}_{ψ}. We notice that $\phi \in \mathcal{L}_{\delta}$ implies $g^{\beta}_{\psi}(\phi) \in \mathcal{L}_{\gamma+\delta}$. The translation $\mathcal{S}^{\beta}_{\psi}$ is thereby defined as: $\mathcal{S}^{\beta}_{\psi}(R\vec{x})$ is $R\vec{x}$ for each atomic formula of $\mathcal{L}_0$, $\mathcal{S}^{\beta}_{\psi}(T_{\delta}x)$ is $T_{\gamma+\delta}\big(g^{\beta}_{\psi}(x)\big)$, $\mathcal{S}^{\beta}_{\psi}(Px)$ is $\psi(x)$, and $\mathcal{S}^{\beta}_{\psi}$ commutes with connectives and quantifiers.

Then, we can straightforwardly show that $\mathsf{RT}_{<\gamma+\beta}[\![\mathsf{B}(P)]\!] \vdash \mathcal{S}^{\beta}_{\psi}(\theta_i)$ for all $i < m$. Since $\theta_j(P) \in \mathcal{L}_0(P)$ and $\mathcal{S}^{\beta}_{\psi}$ keeps the $\mathcal{L}_0$-part unchanged, we have $\mathcal{S}^{\beta}_{\psi}(\theta_j) = \theta_j(\hat{\psi})$ and thus $\mathsf{RT}_{<\gamma+\beta}[\![\mathsf{B}(P)]\!] \vdash \theta_m$. Let $\mathcal{D}_1$ be this derivation of θ_m in $\mathsf{RT}_{<\gamma+\beta}[\![\mathsf{B}(P)]\!]$. Set $\mathcal{D}' := \langle \theta_0, \ldots, \theta_{m-1} \rangle * \mathcal{D}_1 * \langle \theta_{m+1}, \ldots, \theta_n \rangle$, where $*$ denotes the concatenation of sequences. Then $\mathcal{D}'$ is a derivation of θ_n in $(\mathsf{RT}_{<\alpha+\alpha})^* \restriction_k$. $\square$

Lemma 3.5.5. Let α be a limit principal ordinal, i.e., $\alpha = \omega^{\xi}$ for some $\xi > 0$. Then, $(\mathsf{RT}_{<\alpha})^*[\![\mathsf{B}]\!]$ and $\mathsf{RT}_{<\alpha}[\![\mathsf{B}(P)]\!]$ are identical (as systems).

Proof. Suppose $(\mathsf{RT}_{<\alpha})^* \vdash \phi$. Since α is limit, there exist $\beta < \alpha$ and $k \in \mathbb{N}$ such that $\phi \in \mathcal{L}_{\beta}$ and $(\mathsf{RT}_{<\beta})^* \restriction_k \vdash \phi$. Then, it follows from the last lemma that $(\mathsf{RT}_{<\beta \cdot 2^k})^* \restriction_0 \vdash \phi$. Since α is principal, we have $\mathsf{RT}_{<\alpha}[\![\mathsf{B}(P)]\!] \vdash \phi$. $\square$

Theorem 3.5.6. $(\mathsf{RT}_{<\alpha})^* \restriction_k$ is proof-theoretically reducible to $\mathsf{RT}_{<\alpha \cdot 2^k}$ for $\mathcal{L}_0$-formulae. Hence, in particular, $\mathsf{TC}^* \leq \mathsf{RT}_{<\omega}$ $[\mathcal{L}_0]$.

Proof. Suppose $(\mathsf{RT}_{<\alpha})^* \restriction_k \vdash \phi$ for $\mathcal{L}_0$-formula ϕ. By Lemma 3.5.4, we have $\mathsf{RT}_{<\alpha \cdot 2^k}[\![\mathsf{B}(P)]\!] \vdash \phi$ by some derivation $\mathcal{D} = \langle \theta_0, \ldots, \theta_n \rangle$ (θ_n is equal to ϕ). Let $\beta \leq \alpha \cdot 2^k$ be an ordinal such that

$\theta_0, \ldots, \theta_n \in \mathcal{L}_\beta$ and let ψ be an arbitrary $\mathcal{L}_0$-formula (e.g., $(x)_0 = (x)_1$). Take a translation $\mathcal{S}^\beta_\psi$ from $\mathcal{L}_\beta(P)$ to $\mathcal{L}_\beta$ in the same way as Lemma 3.5.4. Then, we can show that $\mathsf{RT}_{<\beta}[\![\mathsf{B}]\!] \vdash \mathcal{S}^\beta_\psi(\theta_n)$. Since $\theta_n \in \mathcal{L}_0$ and $\mathcal{S}^\beta_\psi$ keeps the $\mathcal{L}_0$-part unchanged, we finally have $\mathsf{RT}_{<\alpha\cdot 2^k}[\![\mathsf{B}]\!] \vdash \theta_n$. □

Theorem 3.5.7. Let α be a limit principal ordinal. Then, $(\mathsf{RT}_{<\alpha})^*$ is proof-theoretically equivalent to $\mathsf{RT}_{<\alpha}$ for $\mathcal{L}_0$-formulae.

Proof. Suppose $(\mathsf{RT}_{<\alpha})^* \vdash \phi$ for $\mathcal{L}_0$-formula ϕ. Since α is limit, there exist $\beta < \alpha$ and $k \in \mathbb{N}$ such that $(\mathsf{RT}_{<\beta})^* \restriction_k \vdash \phi$. Then, by the last theorem, we have $\mathsf{RT}_{<\beta\cdot 2^k} \vdash \phi$. Since α is principal, we obtain $\mathsf{RT}_{<\alpha} \vdash \phi$. □

In the same way as Corollary 3.4.7, we can show that:

Corollary 3.5.8. If $\alpha \leq \Gamma_0$, then $\mathsf{DT}^*, \mathsf{KF}^* \not\preceq (\mathsf{RT}_{<\alpha})^*$.

Now we can determine the proof-theoretic strength of FS^*.

Theorem 3.5.9. FS^* is proof-theoretically equivalent to $\mathsf{RT}_{<\omega}$. Thus, in particular, $\mathsf{FS}^*[\![\mathsf{PA}]\!] \equiv \mathsf{RA}_{<\omega}$ (and thus $\equiv \mathsf{ACA}_0$+BR; cf. fn.26 and Theorem 3.5.6).

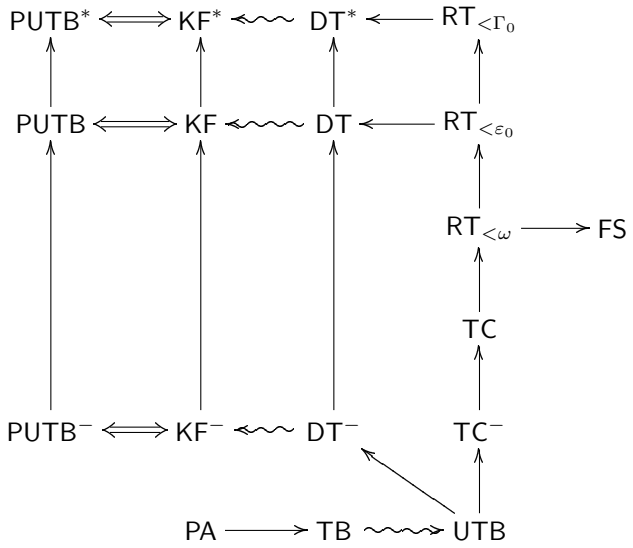
Proof. Halbach [31, §5] already showed that $\mathsf{RT}_{<\omega}$ is truth-definable in FS, and thus $(\mathsf{RT}_{<\omega})^*$ is truth-definable in FS^* by Lemma 3.2.5.

For the converse, Halbach [31, §5] also gave a truth-definition g_n of FS_n in $\mathsf{RT}_{<2n}$, where FS_n is obtained from FS by restricting the numbers of applications of NEC and CONEC in each derivation to at most $(n-1)$-times respectively. It is observed from his proof that if $(\mathsf{FS}_n)^* \vdash \phi$ then $(\mathsf{RT}_{<2n})^* \vdash g_n(\phi)$; for, it can be shown in the same way as [31] (the extra cases for P-Subst are obvious) that $(\mathsf{FS}_i)^* \vdash \phi$ implies $(\mathsf{RT}_{<2n})^* \vdash g_i(\phi) \wedge \cdots \wedge g_{2n-i}(\phi)$ for each $i \leq n$. This implies that $\mathsf{FS}^* \leq (\mathsf{RT}_{<\omega})^*$ and thus $\mathsf{RT}_{<\omega}$ by Theorem 3.5.7.[16] □

[16] Halbach's truth-definitions were for $\mathsf{B} = \mathsf{PA}$, but they can be straightforwardly generalized for an arbitrary B which meets our conditions for base systems.

3.6 Concluding Remarks of Chapter 3

The results on relative truth definability obtained so far are partially summarized by the following diagram: for $\mathsf{B} = \mathsf{PA}$,[17]



In this diagram, for each system Q and S, '$\mathsf{Q} \leftarrow \mathsf{S}$' means $\mathsf{Q} \succ \mathsf{S}$, '$\mathsf{Q} \Leftrightarrow \mathsf{S}$' means $\mathsf{Q} \preceq \mathsf{S}$ and $\mathsf{Q} \succeq \mathsf{S}$, and the curly arrows indicate that it is still open whether the converse holds. The systems in the same horizontal line are all proof-theoretically equivalent (for $\mathcal{L}_0$-sentences); but the systems in the 5th and 6th lines are equivalent as well.

Most of the results on relative truth definability so far are one-way and mutual truth definability fails in many cases. This is what we expected, since we introduced the notion in order to distinguish

[17] As a matter of fact, the diagram applies to an arbitrary base system B (subject to the conditions (a)-(d) in §2.1). However, in order to establish this, we need to show $\mathsf{KF}^* \not\preceq \mathsf{KF}$ and $\mathsf{DT}^* \not\preceq \mathsf{DT}$ for arbitrary B in general, but we have not proved them in the present paper. Its proof needs extra technical arguments which go beyond the scope of the present paper. I only give a sketch of the proof here. We first formulate KF (or DT) in a certain suitable semi-formal system with ω-rule. Then, we can show that if KF (DT) $\vdash \phi$ for an arithmetical ϕ then ϕ is provable in the semi-formal system by a derivation with the length $< \varphi_{\varepsilon_0}0$ which contains no application of cut to formulae with T (but may contain cuts to arithmetical formulae). Since KF^* (DT^*) can prove $\mathrm{TI}_{\mathcal{L}_T}(\varphi_{\varepsilon_0}0)$, it can model those infinitary derivations of arithmetical formulae by a formula Tx: i.e., KF^* (DT^*) can prove that if an arithmetical ϕ is provable in KF (DT) then $T\ulcorner\phi\urcorner$ holds. Thus, KF^* (DT^*) can prove the consistency of KF (DT, resp.).

the systems of truth which have the same truth-free theorems but apparently embody different views of truth. Concluding the present paper, we make some comments on the formal results obtained so far.

First, it is observed that ramified truth systems and type-free truth systems are clearly differentiated in view of relative truth definability. Type-free truth systems are not relatively truth definable in ramified truth systems even with much stronger proof-theoretic strength; e.g., $\mathsf{KF}^{-} \not\preceq \mathsf{RT}_{<\varepsilon_0}$.

Second, relative truth definability can distinguish disquotational systems like UTB and compositional systems like TC^{-}. However, we have also seen that an iterative compositional system KF and a disquotational system PUTB are equivalent in view of relative truth definability. On the one hand, this fact may be regarded as a defect of relative truth definability as a tool for comparing conceptual aspects of systems of truth, since KF and PUTB look different in some respects. For example, although PUTB can define the truth of KF, it can neither prove the compositionality nor iterativity of its own truth predicate in contrast to KF: i.e., PUTB does not prove statements like

$$x, y \in \mathrm{St}_{\mathcal{L}_T} \to \big[T(x \dot{\vee} y) \leftrightarrow Tx \vee Ty\big] \quad \text{or} \quad x \in \mathrm{CT} \to \big[T\dot{\neg}T\dot{x} \leftrightarrow T\dot{\neg}\mathrm{val}(x)\big].$$

For another example, PUTB can be consistently closed by NEC and CONEC, whereas it is not the case for KF; indeed, the resulting system is even conservative over PUTB for truth-free theorems, and, furthermore, the same holds even when we add uniform versions of NEC and CONEC, in which the premises and conclusions of the rules can contain parameters. On the other hand, their mutual truth definability may still be interpreted to indicate that KF and PUTB are indeed 'conceptually equivalent' despite these 'superficial' differences.

Third, it is suggested that iterative compositional systems of truth with weaker evaluation schemata are 'reducible' to those with stronger schemata in a stronger sense than mere relative interpretability, proof-theoretic reducibility and others. But it is still open whether this reduction is

one-way. The solution to the Open Problem 1 would answer the question whether iterative compositional systems of truth with different schemata indeed give different or incompatible conceptions of truth.

Fourth, as we have seen in §5, relative truth definability has useful applications. In particular, relative truth definability makes it easier to compare the schematic reflective closures of systems and mutual truth definability gives more information on model-theoretic features of systems of truth, since it entails conservativity in the semantic sense.

Chapter 4

The Inner Theory of Truth Systems

4.1 The Notion of Inner Theory

Let us start by repeating some historical issues to explain and clarify the motivation of the present chapter. Tarski's theory of truth [82] showed us how the notoriously difficult concept of truth could be formalized in the modern formal framework. As a consequence of his theory, he suggested a negative thesis about formalized truth: i.e., the thesis that no language can possess its own truth predicate and it could be contained only in other essentially richer languages. This thesis naturally leads to the 'orthodox approach' or the 'hierarchical view' of truth [50].

It took four decades since Tarski before a new proposal of formalizing another conception of truth was submitted. Among others (e.g., [56] and [84]), Kripke [50] challenged the Tarskian hierarchical view of truth and presented a method for formalizing a new conception of truth in which the truth predicate is properly applicable to sentences containing the truth predicate itself.

Kripke gave a model-theoretic semantics of a language which contain its own truth predicate and thus his theory counts a semantical theory of truth. Soon after that , an axiomatic treatment of self-applicable truth was presented by Feferman [9, 17]; his system is called Kripke-Feferman system KF today. Since KF, various axiomatic systems of self-applicable truth have been presented and much work has been done on proof-theoretic the analysis of those systems. As I have discussed in Chapter 1, the self-applicable character of the notion of truth can be represented in no definitional theories of truth and one of the major advantages of the non-definitional approach consists in the fact they can properly deal with it. The present chapter particularly focuses on axiomatic systems of self-applicable truth and proposes one means of comparison among them.

Since a system of self-applicable truth has its own truth predicate, it is naturally motivated to distinguish and compare a statement 'A' and the assertion of its truthhood 'the statement A is true' within the same system; both types of statements are expressible in such a system of self-applicable truth. This distinction and comparison were first explicitly made by Reinhardt [70, 71]. He studied the class of the statements σ which are *asserted to be true* or *provably true* in KF, i.e., $\{\sigma \in \mathcal{L}_T \mid \mathsf{KF} \vdash T\ulcorner\sigma\urcorner\}$. This class is called the *inner theory* of KF. This approach of Reinhardt and the notion of inner theory can be naturally generalized to any other system Q of self-applicable truth: i.e., the inner theory of Q is $\{\sigma \mid \sigma \text{ is a sentence of } \mathsf{Q} \text{ and } \mathsf{Q} \vdash T\ulcorner\sigma\urcorner\}$. Let us denote it by $\mathcal{I}(\mathsf{Q})$. In contrast against the inner theory, we call the set of theorems of Q the *outer theory* of Q and denote it by $\mathcal{O}(\mathsf{Q})$: i.e., $\mathcal{O}(\mathsf{Q}) = \{\sigma \mid \mathsf{Q} \vdash \sigma\}$. As we will argue in the following, the notion of inner theory has philosophical and mathematical significance in some respects. The objective of the present chapter is to investigate the inner theories of various systems of self-applicable truth and correlate these systems in terms of their inner theories.

Now let us look at Reinhardt's original motivation of his study of inner theories. Reinhardt interpreted KF as a system of *meaningful applicability*, the notion of which he ascribed to Gödel [25]. This idea and its motivation are briefly summarized by the following passage: '[T]he paradoxes can

perhaps be resolved by noting that some properties and relations ... may not always be meaningfully applicable ... [70, p.227]'; in other words, Reinhardt suggests that the paradoxes such as Russell's paradox and the Liar paradox may be resolved by restricting the application of the principles which leads to these paradoxes (i.e., the full comprehension axiom of set theory and the full T-schema in truth theory) only to a certain domain the objects of which those principles are *meaningfully applicable* to. First, Reinhardt argued that a sentence σ is meaningfully applicable or *significant* (they are equivalent notions for Reinhardt) if and only if σ has determinate truth value, i.e., truth or falsity.[1] Then, Reinhardt proposed to use KF as a tool of producing true (and thus significant) statements and make the following thesis:

(R1) If A is a sentence such that $\mathsf{KF} \vdash T[A]$, then A is true.

(R1) If A is a sentence such that $\mathsf{KF} \vdash T[A] \vee F[A]$, then A is significant.

Hence, for Reinhardt, the inner theory of KF comprises true and significant sentences. Reinhardt makes an analogy between his project and Hilbert's programme; in this analogy of his, $\mathcal{O}(\mathsf{KF})$ and $\mathcal{I}(\mathsf{KF})$ respectively corresponds to Hilbert's 'ideal' mathematics and 'real' (or 'contentual') mathematics. Then, Reinhardt raised two questions of KF ([70, p.239]):

(1) Is there a natural axiomatization of $\mathcal{I}(\mathsf{KF})$ which consists of only significant sentences?

(2) If $\mathsf{KF} \vdash T^{\ulcorner}\sigma^{\urcorner}$, is there a proof $\mathcal{D} = \langle \theta_0, \ldots, \theta_n \rangle$ of σ such that $\mathsf{KF} \vdash T^{\ulcorner}\theta_i^{\urcorner}$ for each $i \leq n$?

The question (1) is important for Reinhardt since a positive answer to it implies that we have natural axiomatization of the set of true statements by true sentences. A positive answer to (2) 'justifies the use of insignificant sentences' as a mere tool for deducing true sentences 'entirely within the framework of the significant sentences'. Here we can again see an analogy between Reinhardt's project and Hilbert's programme; (2) corresponds to the conservativity requirement of Hilbert's

[1] This view is also taken by Feferman and motivated his theory DT, though Feferman's motivation is probably independent of Reinhardt.

programme. Consequently, since we can show that the desired truth-theoretic principles such as the T-schema and compositional axioms (or inductive clauses) hold of significant sentences within KF, KF would then be regarded as a fairly satisfactory system of truth for Reinhardt.

Of course, the answer to (1) depends on what 'natural' means there; but, at any rate, it is easily seen that the answer to (1) would be negative if the required axiomatization need to be based on classical logic, since $\lambda \vee \neg\lambda \notin \mathcal{I}(\mathsf{KF})$ where λ is the liar sentence. To make the matter worse, (2) was already negatively answered by Halbach and Horsten [38]. However, at the time of Reinhardt's writing [70] and [71], no alternative axiomatic system of self-applicable truth to KF was known, whereas we now have much more variety of such systems. Thus, we may propose the thesis like (**R1**) and (**R2**) with respect to other systems of self-applicable truth and then ask similar questions to (1) and (2) for them. We have already seen that the three systems of symmetric truths, i.e., FS, POSKF + NEC + CONEC and PUTB + NEC + CONEC (as well as their variants), and they all meet both (1) and (2); in addition, more systems such as KF + Comp meet (2) although they fail to meet (1). It is to be noted, however, that the positive answers to (1) and (2) were of importance to Reinhard because he held the theses (**R1**) and (**R2**) particularly with respect to KF. Hence, it is still debatable whether other systems can be a substitute for KF for Reinhardt's project. At any rate, Reinhardt's project can be generalized and has not yet been put an end.

The notion of inner theory thus arose from Reinhardt's specific project and philosophical motivation. However, besides Reinhardt's project, it seems to be philosophically and mathematically significant in other respects. In what follows, we will explain some of them.

From the mathematical point of view, the notion of inner theory could be a useful tool for comparing systems of self-applicable truth. As in the case of relative truth definability, comparison by inner theory brings more information about systems of self-applicable truth than comparison by the mere derivability of truth-free theorems. I repeatedly emphasize that the proof-theoretic analyses so far given only focus on the truth-free part of systems of truth. However, we often

compare two subsystems of second-order arithmetic by their derivability of a certain larger class of formulae; this type of comparison is often found in the subject of reverse mathematics (cf. [77]). For example, Σ^1_1-DC and $(\Pi^0_1\text{-CA})_{<\varepsilon_0}$, which are both proof-theoretically equivalent to KF, prove the same Π^1_2-sentences but the former has properly more Π^1_3 theorems than the latter. In view of such a practice in mathematical logic, it seems natural to consider and compare by larger classes of theorems of truth systems than that of arithmetical ones.

One natural generalization of the existing approach is perhaps to expand an arithmetical hierarchy by incorporating the truth predicate and then to compare systems in the light of this new hierarchy. We can define classes Π^T_0 and Σ^T_0 of formulae as the least set containing all arithmetical atomics and those of the form $Tt(\vec{x})$ closed under Boolean combinations and bounded quantifications; then Σ^T_n and Π^T_n can be defined in parallel manners to the ordinary arithmetical hierarchy. This direction should be of some interest. However, in many cases, systems of truth are incomparable even for Π^T_1 or Σ^T_1 class: that is, the class of Π^T_1 (Σ^T_1) theorems of one system neither includes nor is included in that of another system in many cases. As we shall see in §3, as far as the systems considered in this chapter concern, they are all incomparable by means of thus defined expanded arithmetical hierarchy except for the trivial cases in which one is a subtheory of another.

In contrast to the comparison of the above expanded arithmetical hierarchy, many theories are suitably comparable to one another by means of their inner theories. Hence, I suggest the comparison of inner theories as a better behaved tool to compare systems of self-applicable truth. I also notice that, since every system considered in the present thesis derives T-biconditionals for arithmetical (truth-free) formulae, the inner theory comparison is properly finer grained than the mere comparison by arithmetical (truth-free) consequences.

Comparison of systems of self-applicable truth by their inner theories also has a philosophical motivation. Different systems of truth embody different views or conceptions of truth. However, as we have extensively discussed in the last chapter, comparison only by the arithmetical part

often bring no difference among systems of truth based on different views on truth. Now, let us consider KF and WKF for instance. They are proof-theoretically equivalent, but we probably have the intuition (or at least the feeling) that weak Kleene logic is weaker than strong Kleene logic in some sense (judging not simply from their names 'weak Kleene' and 'strong Kleene' but from their truth tables) and thus WKF should be weaker than KF in some sense. This intuition is correctly reflected by their inner theories; we will show that $\mathcal{I}(\mathsf{DT}) \subsetneq \mathcal{I}(\mathsf{KF})$, and thus WKF is indeed 'weaker' than KF in view of their inner theories.

The present chapter is basically a technical one but motivated by the aforementioned rather philosophical considerations. In what follows, we always assume that the base system B is PA; as I have explained, we no longer need to discuss within the former general setting after obtaining some results concerning relative truth definability in Chapter 3.

4.2 Systems of Finitely Iterated Self-Applicable Truth

For the discussion that follows, we technically need to introduce the systems of finitely iterated self-applicable truth; they will be used for the analysis of the inner theories of schematic reflective closures of KF, WKF and FKF (we will later consider the systems of transfinitely iterated self-applicable truth in the next chapter).

Definition 4.2.1. For each natural number $n > 0$, we introduce a primitive recursive extension KF_n of KF. The language of KF_n consists of $\mathcal{L}_n = \mathcal{L}_0 \cup \{T_k\}_{1 \leq k \leq n}$, where T_k's are new unary predicate. We set KF_0 to be PA. For $n > 0$, the system KF_n over $\mathcal{L}_n$ consists of KF_{n-1}, the full induction for $\mathcal{L}_n$ and the following axioms for the predicate T_n:

K1$_n$ $\forall x, y \in \mathrm{CT}\big[\big(T_n(x \dot{=} y) \leftrightarrow x^\circ = y^\circ\big) \wedge \big(F_n(x \dot{=} y) \leftrightarrow x^\circ \neq y^\circ\big)\big]$.

K2$_n$ $\forall \mathrm{CT}(x)\big[\big(T_n(\dot{T}_n x) \leftrightarrow T_n x^\circ\big) \wedge \big(F_n(\dot{T}_n x) \leftrightarrow F_n x^\circ\big)$.

K3$_n$ $\forall x \in \mathrm{St}_n[T_n(\dot{\neg}\dot{\neg} x) \leftrightarrow T_n x]$

$\mathbf{K4}_n$ $\forall x, y \in \mathrm{St}_n\big[\big(T_n(x \dot{\wedge} y) \leftrightarrow (T_n x \wedge T_n y)\big) \wedge \big(F_n(x \dot{\wedge} y) \leftrightarrow (F_n \vee F_n y)\big)\big]$

$\mathbf{K5}_n$ $\forall x, y \in \mathrm{St}_n\big[\big(T_n(x \dot{\rightarrow} y) \leftrightarrow (F_n x \vee T_n y)\big) \wedge \big(F_n(x \dot{\rightarrow} y) \leftrightarrow (T_n \wedge F_n y)\big)\big]$

$\mathbf{K6}_n$ $\forall x \forall y\big[\mathrm{St}_n(\dot{\forall} x.y) \rightarrow \big(T_n(\dot{\forall} x.y) \leftrightarrow \forall z T_n y(z)\big) \wedge \big(F_n(\dot{\forall} x.y) \leftrightarrow \exists z F_n y(z)\big)\big]$

$\mathbf{K7}_n$ $\forall x \in \mathrm{CT}\big[\big(T_n(\dot{T}_m x) \leftrightarrow T_m x^\circ\big) \wedge \big(F_n(\dot{T}_m x) \leftrightarrow \neg T_m x^\circ\big)\big]$, for each $m < n$.

Then, we can identify KF and KF_1. We also define the system $\mathsf{KF}_{<\omega}$ as $\bigcup_n \mathsf{KF}_n$.

WKF_n and $\mathsf{WKF}_{<\omega}$ (FKF^n and $\mathsf{WKF}^{<\omega}$) are analogously defined.

Lemma 4.2.2. Let $n > m$. For every $\mathcal{L}_m$-formula $\phi(\vec{x})$, $\mathsf{KF}_n \vdash \forall \vec{x}[T_n(\ulcorner \phi(\dot{\vec{x}}) \urcorner) \leftrightarrow \phi(\vec{x})]$. WKF_n and FKF_n derive the same.

Corollary 4.2.3. Let $n > m$ and $\phi(\vec{x})$ be an $\mathcal{L}_m$-formula. If $\mathsf{KF}_m \vdash \phi(\vec{x})$, then $\mathsf{KF}_n \vdash T_n(\ulcorner \phi(\dot{\vec{x}}) \urcorner)$. The same also holds in the cases of the iterated WKF and FKF.

Lemma 4.2.4. Let $n \geq m \geq l$. We have $\mathsf{KF}_n \vdash \forall x \in \mathrm{St}_m(T_m x \leftrightarrow T_l x)$. WKF_n and FKF_n derive the same.

For the reader's convenience, we repeat the definition of $\mathsf{Q}^* \restriction_n$ made in §5 of the last chapter: given $n \in \mathbb{N}$ and a system Q, the system $\mathsf{Q}^* \restriction_n$ is obtained from the schematic reflective closure Q^* by restricting the number of application of the P-Subst rule to at most n-times. Recall that the language of $(\mathsf{KF}_n)^*$ is $\mathcal{L}_n(P)$ which contains a new (vacuous) predicate P. In the rest of the present section, we assume that each derivation of $(\mathsf{KF}_n)^*$ is given in Hilbert style in which the inference rules are either Modus Ponens or P-Subst.

The proof of the next lemma is parallel to that of of Lemma 3.5.4 .

Lemma 4.2.5. $(\mathsf{KF}_n)^* \restriction_{k+1}$ is a subtheory of $(\mathsf{KF}_{2n})^* \restriction_k$.

Proof. Given a derivation $\mathcal{D} = \langle \theta_0, \ldots, \theta_l \rangle$ in $(\mathsf{KF}_n)^* \restriction_{k+1}$, let $\mathcal{D}_0 = \langle \theta_0, \ldots, \theta_{l'} \rangle$ be the initial segment of $\mathcal{D}$ which ends with the first application of P-Subst $\frac{\theta_j(P)}{\theta_j(\hat{\psi})}$ for some $j < l'$; thus, $l' \leq l$ and $\theta_{l'} \equiv \theta_j(\hat{\psi})$.

Let $\psi \in \mathcal{L}_m(P)$ $(m \leq n)$. We define a translation $\mathcal{S}^n_\psi$ from $\mathcal{L}_n(P)$ to $\mathcal{L}_{m+n}(P)$, which 'lifts up' the level of truth predicates in each θ_i $(i < m)$ so that ψ can be treated as if it were at the lowest level. First we preliminarily define a primitive recursive function g^n_ψ (by the primitive recursion theorem) by:

$$g^n_\psi(a) := \begin{cases} a & \text{if } a \in \mathrm{AtSt}_0 \\ \ulcorner \psi(t) \urcorner & \text{if } a \text{ is } \ulcorner P(t) \urcorner \\ \ulcorner T_{m+i}\big(g^\beta_\psi(t)\big) \urcorner & \text{if } a \text{ is } \ulcorner T_i t \urcorner \ (1 \leq i \leq n) \\ \dot{\neg} g^n_\psi(b) & \text{if } a = \dot{\neg} b \\ g^n_\psi(b) \dot{\wedge} g^n_\psi(c) \quad \big(g^n_\psi(b) \dot{\to} g^n_\psi(c), \text{resp.}\big) & \text{if } a = b \dot{\wedge} c \text{ (or } a = b \dot{\to} c) \\ \dot{\forall} z. g^n_\psi(b) & \text{if } a = \dot{\forall} z. b. \end{cases}$$

$\mathcal{S}^n_\psi$ is thereby defined as: $\mathcal{S}^n_\psi(R\vec{x})$ is $R\vec{x}$ for each atomic formula of $\mathcal{L}_0$, $\mathcal{S}^n_\psi(T_i x)$ is $T_{m+i}\big(g^n_\psi(x)\big)$, $\mathcal{S}^n_\psi(Px)$ is $\psi(x)$, and $\mathcal{S}^n_\psi$ commutes with the connectives and quantifiers.

We can straightforwardly show that $(\mathsf{KF}_{m+n})^* \restriction_0 \vdash \mathcal{S}^n_\psi(\theta_q)$ for each $q < l'$. We thus obtain $(\mathsf{KF}_{m+n})^* \restriction_0 \vdash \mathcal{S}^n_\psi\big(\theta_j(P)\big)$. Since $\mathcal{S}^n_\psi$ preserves connectives and quantifiers and keeps the $\mathcal{L}_0$-part unchanged, we have $\mathcal{S}^n_\psi\big(\theta_j(P)\big) \equiv \theta_j(\hat{\psi})$ $(\equiv \theta_{l'})$. Let $\mathcal{D}_1$ be this derivation of $\theta_{l'}$ in $(\mathsf{KF}_{m+n})^* \restriction_0$. Set $\mathcal{D}' := \langle \theta_0, \ldots, \theta_{l'-1} \rangle * \mathcal{D}_1 * \langle \theta_{l'+1}, \ldots, \theta_l \rangle$, where $*$ denotes the concatenation of sequences. Then $\mathcal{D}'$ is a derivation of θ_l in $(\mathsf{KF}_{m+n})^* \restriction_k$. □

Corollary 4.2.6. $(\mathsf{KF}_n)^* \restriction_k$ is a subtheory of $(\mathsf{KF}_{2^k \cdot n})^* \restriction_0$.

Corollary 4.2.7. $(\mathsf{KF}_n)^*$ is a subtheory of $(\mathsf{KF}_{<\omega})^* \restriction_0$.

Corollary 4.2.8. $(\mathsf{KF}_{<\omega})^*$ is a subtheory of $(\mathsf{KF}_{<\omega})^* \restriction_0$.

Lemma 4.2.9. $(\mathsf{KF}_n)^*$ is proof-theoretically reducible to $\mathsf{KF}_{<\omega}$ for $\mathcal{L}_0$-formulae.

Proof. Suppose $(\mathsf{KF}_n)^* \vdash \phi$ for $\phi \in \mathcal{L}_0$. For some $k \in \mathbb{N}$, we have $(\mathsf{KF}_k)^* \restriction_0 \vdash \phi$ by some derivation $\mathcal{D} = \langle \theta_0, \ldots, \theta_m \rangle$ (θ_m is equal to ϕ). Let ψ be an arbitrary $\mathcal{L}_0$-formula (e.g., $(x)_0 = (x)_1$). Take a translation $\mathcal{S}^k_\psi$ from $\mathcal{L}_k(P)$ to $\mathcal{L}_k$ in the same way as the last lemma. Then, we can show that

$\mathsf{KF}_k \vdash \mathcal{S}^k_\psi(\theta_m)$. Since $\theta_m \in \mathcal{L}_0$ and $\mathcal{S}^k_\psi$ keeps the $\mathcal{L}_0$-part unchanged, we finally have $\mathsf{KF}_k \vdash \theta_m$. $\square$

One might expect that lemma 4.2.9 would hold even for the case where $\phi \in \mathcal{L}_n$ and it would hold that $(\mathsf{KF}_n)^* \leq \mathsf{KF}_{<\omega}$ $[\mathcal{L}_n]$. However, it is not the case since $T^\ulcorner P\dot{x} \vee \neg P\dot{x}^\urcorner$ is provable in $(\mathsf{KF}_m)^*$ but not in $\mathsf{KF}_{<\omega}$.

Next, we turn to the schematic reflective closures with Cons. For a given $n \in \mathbb{N}$, let Cons_n and Comp_n respectively denote the axioms of consistency and completeness with respect to the n-iterated truth: i.e.,

$$\mathrm{Cons}_n \equiv \forall x \in \mathrm{St}_n \neg[T_n(x) \wedge T_n(\dot{\neg} x)] \qquad \text{and} \qquad \mathrm{Comp}_n \equiv \forall x \in \mathrm{St}_n[T_n(x) \vee T_n(\dot{\neg} x)];$$

we define $\mathrm{Cons}_{<\omega} := \bigcup_{n\in\mathbb{N}} \mathrm{Cons}_n$ and $\mathrm{Comp}_{<\omega} := \bigcup_{n\in\mathbb{N}} \mathrm{Comp}_n$.

Proposition 4.2.10. Let $m \leq n$. Then, $\mathsf{KF}_n \vdash (\mathrm{Cons}_n \to \mathrm{Cons}_m) \wedge (\mathrm{Comp}_n \to \mathrm{Comp}_m)$. The same for WKF_n and FKF_n.

Lemma 4.2.11. $\mathsf{KF}_n + \mathrm{Cons}_n$ and $\mathsf{KF}_n + \mathrm{Comp}_n$ are mutually truth definable. As in the case of non-iterated truth, this doesn't hold for WKF_n and FKF_n.

Proof. We define in PA (by the recursion theorem or diagonalization) a function f by:

$$f(x) := \begin{cases} x & \text{if } x \in \mathrm{AtFml}_0 \\ \ulcorner \neg F_m(fy) \urcorner \ \big(\text{i.e., } \mathrm{sb}(\ulcorner \neg F_m f(\cdot) \urcorner; y)\big) & \text{if } x = \dot{T}_m y \text{ for } y \in \mathrm{Tm} \text{ and } m \leq n \\ \dot{\neg} f(y) \ \ (f(y) \dot{\wedge} f(z), \text{ resp.}) & \text{if } x = \dot{\neg} y \text{ (or } x = y \dot{\wedge} z) \\ \dot{\forall} z.f(y) & \text{if } x = \dot{\forall} z.y; \end{cases}$$

otherwise, $f(x) = 0$. The truth definition $\mathcal{T}$ for both claims is given by $T_z^{\prec} x \mapsto \neg F_z f(x)$. We only

demonstrate that $\mathbf{K2}_n$ is preserved by $\mathcal{T}$: for $x \in \mathrm{CT}$ and $m < n$,

$$\mathcal{T}(T_n \dot{T}_m x) \equiv \neg F_n \ulcorner \neg F_m(fx) \urcorner \leftrightarrow \neg T_n \ulcorner F_m(fx) \urcorner \leftrightarrow \neg F_m f(x^\circ) \equiv \mathcal{T}(T_m x^\circ)$$

$$\mathcal{T}(F_n \dot{T}_m x) \equiv \neg F_n \dot{\neg} \ulcorner \neg F_m(fx) \urcorner \leftrightarrow \neg F_n \ulcorner F_m(fx) \urcorner \leftrightarrow \neg\neg F_n f(x^\circ) \equiv \mathcal{T}(\neg T_m x^\circ).$$

The other cases are straightforward. □

Lemma 4.2.12. $(\mathsf{KF}_n + \mathrm{Cons}_n)^* \restriction_{k+1}$ is a subtheory of $(\mathsf{KF}_{2n} + \mathrm{Cons}_{2n})^* \restriction_k$.

Corollary 4.2.13. $(\mathsf{KF}_n + \mathrm{Cons}_n)^* \restriction_k$ is a subtheory of $(\mathsf{KF}_{2^k \cdot n} + \mathrm{Cons}_{2^k \cdot n})^* \restriction_0$.

Corollary 4.2.14. $(\mathsf{KF}_n + \mathrm{Cons}_n)^*$ is a subtheory of $(\mathsf{KF}_{<\omega} + \mathrm{Cons}_{<\omega})^* \restriction_0$.

Corollary 4.2.15. $(\mathsf{KF}_{<\omega} + \mathrm{Cons}_{<\omega})^*$ is a subtheory of $(\mathsf{KF}_{<\omega} + \mathrm{Cons}_{<\omega})^* \restriction_0$.

Lemma 4.2.16. $(\mathsf{KF}_n + \mathrm{Cons}_n)^*$ is proof-theoretically reducible to $\mathsf{KF}_{<\omega} + \mathrm{Cons}_{<\omega}$ for $\mathcal{L}_0$-formulae.

4.3 Semantics of iterative compositional theories

In the present section, we generalize the semantics of KF^- developed by Cantini [5] to the other systems WKF^-, FKF^- and PUTB^-. It will be used to determine how their inner theories of truth systems are correlated.

Let M be an arbitrary model of PA, and fix this M throughout this section. $|M|$ denote the domain of M. Given a formula ϕ and a function f of $\mathcal{L}$, ϕ^M and f^M denote the interpretation of ϕ and f in M respectively. Given a set $\mathcal{C}$ of subsets of $|M|$ (i.e., $C \subset \mathcal{P}(|M|)$), $\sup \mathcal{C}$ and $\inf \mathcal{C}$ denote $\bigcup \mathcal{C}$ ($\subset |M|$) and $\bigcap \mathcal{C}$ ($\subset |M|$) respectively. For each $Y \subset |M|$, we write $-Y$ for $|M| \setminus Y$. For an $\mathcal{L}_T$-formula ϕ and $\vec{a} \in |M|$, we write $X \models \phi(\vec{a})$ iff $(M, X) \models \phi(\vec{a})$, where T is interpreted by X.

Definition 4.3.1. **(1)** $X \subset |M|$ is regular iff, for any $a \in \mathrm{For}^M$ and $b \in \mathrm{CT}^M$,

$$\mathrm{sb}^M(a,b) \in X \Leftrightarrow \mathrm{sb}^M\big(a, \mathrm{nm}(\mathrm{val}(b))\big) \in X.$$

We denote the set of all regular $X \subset |M|$ by REG_M. Note that $X \in \mathrm{REG}_M$ assures that X is a model of the axiom K0.

(2) Given $X \in \mathrm{REG}_M$,

- we say that X is consistent iff there is no $a \in |M| \cap \mathrm{St}^M_{\mathcal{L}_T}$ such that both a and $\dot{\neg}^M a$ are in X, and CONS_M denotes the set of all consistent $X \in \mathrm{REG}_M$;
- we say that X is complete iff either a or $\dot{\neg}^M a$ is in X for every $a \in |M| \cap \mathrm{St}^M_{\mathcal{L}_T}$, and COMP_M denotes the set of all complete $X \in \mathrm{REG}_M$

Definition 4.3.2. We will define operators **S**, **F** and **W** from REG_M to REG_M according to Strong Kleene, Feferman, and Weak Kleene schemata, respectively.

(I) $a \in \mathbf{W}(X)$ iff $a \in \mathrm{St}^M_{\mathcal{L}_T}$ and

(i) $M \models a \doteq \dot{R}(\vec{b}) \wedge \mathrm{CT}(\vec{b}) \wedge R(\vec{b}^\circ)$, or

(ii) $M \models a \doteq \dot{\neg}\dot{R}(\vec{b}) \wedge \mathrm{CT}(\vec{b}) \wedge \neg R(\vec{b}^\circ)$, or

(iii) $M \models \mathrm{CT}(b) \wedge a \doteq \dot{T}b$ and $X \models T(b^\circ)$, or

(iv) $M \models \mathrm{CT}(b) \wedge a \doteq \dot{\neg}\dot{T}b$ and $X \models F(b^\circ)$, or

(v) $M \models \mathrm{CT}(b) \wedge a \doteq \dot{F}b$ and $X \models F(b^\circ)$, or

(vi) $M \models \mathrm{CT}(b) \wedge a \doteq \dot{\neg}\dot{F}b$ and $X \models T(b^\circ)$, or

(vii) $M \models a \doteq \dot{\neg}\dot{\neg}b$ and $X \models Tb$, or

(viii) $M \models a \doteq b \dot{\wedge} c$ and $X \models Tb \wedge Tc$, or

(ix) $M \models a \doteq \dot{\neg}(b \dot{\wedge} c)$ and $X \models (Fb \wedge Fc) \vee (Tb \wedge Fc) \vee (Fb \dot{\wedge} Tc)$, or

(x) $M \models a = \dot{\forall} z.a$ and $X \models \forall z Ta(z)$, or

(xi) $M \models a = \dot{\neg}(\dot{\forall} z.a)$ and $X \models \big[\forall z\big(Ta(z) \vee Fa(z)\big)\big] \wedge \exists z Fa(z)$.

(II) $a \in \mathbf{F}(X)$ iff $a \in \mathrm{St}^M_{\mathcal{L}_T}$ and either of (i)-(xi) and the following (xii)–(xiii) holds for a:

(xii) $M \models a = b \dot{\rightarrow} c$ and $X \models (Ta \wedge Tc) \vee Fb$

(xiii) $M \models a = \dot{\neg}(b \dot{\rightarrow} c)$ and $X \models (Ta \wedge Fc)$.

(III) $a \in \mathbf{S}(X)$ iff $a \in \mathrm{St}^M_{\mathcal{L}_T}$, and one of (i)-(vii), (xii)-(xiii) and the following holds

(viii') $M \models a = b \dot{\wedge} c$ and $X \models Tb \wedge Tc$ or

(ix') $M \models a = \dot{\neg}(b \dot{\wedge} c)$ and $X \models T\dot{\neg} b \vee T\dot{\neg} b$ or

(x') $M \models \mathrm{Var}(z) \wedge a = \dot{\forall} z.a$ and $X \models \forall z Ta(z)$ or

(xi') $M \models \mathrm{Var}(z) \wedge a = \dot{\neg}\dot{\forall} z.a$ and $X \models \exists z T\dot{\neg} a(z)$.

Proposition 4.3.3. $\mathbf{S}$, $\mathbf{F}$ and $\mathbf{W}$ are operators from REG_M to REG_M.

Proof. The proof for $\mathbf{S}$ is already given in [5]. The other cases are parallel. We remark that the assumption that sb commutes with (the representations of) logical operations is needed to prove this proposition. □

For an operator $\mathbf{G}$ from $\mathcal{P}(|M|)$ to $\mathcal{P}(|M|)$, we call it monotone if $\mathbf{G}(X) \subset \mathbf{G}(Y)$ whenever $X \subset Y$. Then it is observed that $\mathbf{S}$, $\mathbf{F}$ and $\mathbf{W}$ are all monotone.

Definition 4.3.4. Let $\mathbf{G}$ be any monotone operator from $\mathcal{P}(|M|)$ to $\mathcal{P}(|M|)$ and let $X \subset |M|$.

1. For each ordinal α, $X^{\mathbf{G}}_\alpha$ is recursively defined by:

$$X^{\mathbf{G}}_0 = X, \qquad X^{\mathbf{G}}_{\beta+1} = X^{\mathbf{G}}_\beta \cup \mathbf{G}(X_\beta), \qquad X^{\mathbf{G}}_\lambda = \bigcup_{\beta<\lambda} X^{\mathbf{G}}_\beta,$$

where λ is a limit ordinal. Dually, $X^{\alpha}_{\mathbf{G}}$ is defined by

$$X^{0}_{\mathbf{G}} = X, \qquad X^{\beta+1}_{\mathbf{G}} = X^{\beta}_{\mathbf{G}} \cap \mathbf{G}(X^{\beta}_{\mathbf{G}}), \qquad X^{\lambda}_{\mathbf{G}} = \bigcap_{\beta<\lambda} X^{\beta}_{\mathbf{G}}.$$

2. $\mathrm{UP}^{\mathbf{G}}(X) \stackrel{\mathrm{def}}{=} X^{\mathbf{G}}_{\infty} = \bigcup_{\alpha\in On} X^{\mathbf{G}}_{\alpha}$. $\mathrm{DOWN}^{\mathbf{G}}(X) \stackrel{\mathrm{def}}{=} X^{\infty}_{\mathbf{G}} = \bigcap_{\alpha\in On} X^{\alpha}_{\mathbf{G}}$.

3. $X^{\mathbf{G}}_{\mathrm{up}} \stackrel{\mathrm{def}}{=} \mathrm{DOWN}^{\mathbf{G}}(\mathrm{UP}^{\mathbf{G}}(X))$, and $X^{\mathbf{G}}_{\mathrm{down}} \stackrel{\mathrm{def}}{=} \mathrm{UP}^{\mathbf{G}}(\mathrm{DOWN}^{\mathbf{G}}(X))$.

4. $\mathrm{FIX}^{\mathbf{G}}_{M} \stackrel{\mathrm{def}}{=} \{X \in \mathrm{REG}_M \mid \mathbf{G}(X) = X\}$.

Given $X \subset |M|$, we say X is $\mathbf{G}$-dense if $X \subset \mathbf{G}(X)$, and X is $\mathbf{G}$-closed if $\mathbf{G}(X) \subset X$.

Proposition 4.3.5. For any monotone operator $\mathbf{G}\colon \mathcal{P}(|M|) \to \mathcal{P}(|M|)$, the following hold:

(1) $\mathrm{UP}^{\mathbf{G}}(X)$ is the the $\subset$-least $\mathbf{G}$-closed set including X among REG_M, and $\mathrm{DOWN}^{\mathbf{G}}(X)$ is the $\subset$-largest $\mathbf{G}$-dense set included in X among REG_M.

(2) $X^{\mathbf{G}}_{\mathrm{up}}, X^{\mathbf{G}}_{\mathrm{down}} \in \mathrm{FIX}^{\mathbf{G}}_{M}$ and $X^{\mathbf{G}}_{\mathrm{down}} \subset X^{\mathbf{G}}_{\mathrm{up}}$.

(3) $\mathrm{UP}^{\mathbf{G}}(\emptyset) = \emptyset^{\mathbf{G}}_{\mathrm{up}} = \emptyset^{\mathbf{G}}_{\mathrm{down}}$, and $\mathrm{DOWN}^{\mathbf{G}}(|M|) = |M|^{\mathbf{G}}_{\mathrm{down}} = |M|^{\mathbf{G}}_{\mathrm{up}}$. Therefore, $\emptyset^{\mathbf{G}}_{\mathrm{up}}$ is minimal fixed-point and $|M|^{\mathbf{G}}_{\mathrm{down}}$ is maximal fixed-point with respect to $\mathbf{G}$.

(4) If $\mathbf{G}$ preserves regularity (i.e., $\mathbf{G}(X) \in \mathrm{REG}_M$ whenever $X \in \mathrm{REG}_M$), then $X^{\mathbf{G}}_{\alpha}, X^{\alpha}_{\mathbf{G}} \in \mathrm{REG}_M$ for all $X \in \mathrm{REG}_M$.

Proposition 4.3.6. Let $\mathbf{G}$ be $\mathbf{W}$, $\mathbf{F}$ or $\mathbf{S}$. If $X \in \mathrm{CONS}_M$ then $X^{\mathbf{G}}_{\mathrm{down}} \in \mathrm{CONS}_M$, and if $X \in \mathrm{COMP}_M$ then $X^{\mathbf{G}}_{\mathrm{up}} \in \mathrm{COMP}_M$.

Proposition 4.3.7. Let $\mathbf{G}$ be $\mathbf{S}$, $\mathbf{F}$ or $\mathbf{W}$. Then, $\emptyset^{\mathbf{G}}_{\mathrm{up}} \in \mathrm{CONS}_M$ and $|M|^{\mathbf{G}}_{\mathrm{down}} \in \mathrm{COMP}_M$.

Lemma 4.3.8. Let $X \subset |M|$.

(1) $X \in \mathrm{FIX}^{\mathbf{W}}_{M}$ iff $X \models \mathrm{WKF}^{-}$. $X \in \mathrm{FIX}^{\mathbf{W}}_{M} \cap \mathrm{CONS}_M$ iff $X \models \mathrm{WKF}^{-} + \mathrm{Cons}$. $X \in \mathrm{FIX}^{\mathbf{W}}_{M} \cap \mathrm{COMP}_M$ iff $X \models \mathrm{WKF}^{-} + \mathrm{Comp}$.

(2) $X \in \mathrm{FIX}^{\mathbf{F}}_M$ iff $X \models \mathrm{FKF}^-$. $X \in \mathrm{FIX}^{\mathbf{F}}_M \cap \mathrm{CONS}_M$ iff $X \models \mathrm{FKF}^- + \mathrm{Cons}$. $X \in \mathrm{FIX}^{\mathbf{F}}_M \cap \mathrm{COMP}_M$ iff $X \models \mathrm{FKF}^- + \mathrm{Comp}$.

(3) $X \in \mathrm{FIX}^{\mathbf{S}}_M$ iff $X \models \mathrm{KF}^-$. $X \in \mathrm{FIX}^{\mathbf{S}}_M \cap \mathrm{CONS}_M$ iff $X \models \mathrm{KF}^- + \mathrm{Cons}$. $X \in \mathrm{FIX}^{\mathbf{S}}_M \cap \mathrm{COMP}_M$ iff $X \models \mathrm{KF}^- + \mathrm{Comp}$.

(4) If M is standard, (1)-(3) hold even for WKF, FKF and KF respectively.

Definition 4.3.9. We also consider the semantics of PUTB. In the following SPOS_M and SNEG_M denotes the set

$$\{\mathrm{sb}^M_{\vec{x}}(\ulcorner \phi(\vec{x}) \urcorner, \vec{b}) \mid \phi(\vec{x}) \text{ is a (standard) } T\text{-positive } \mathcal{L}_T\text{-formula and } \vec{b} \in \mathrm{CT}^M\}$$

$$\{\mathrm{sb}^M_{\vec{x}}(\ulcorner \phi(\vec{x}) \urcorner, \vec{b}) \mid \phi(\vec{x}) \text{ is a (standard) } T\text{-negative } \mathcal{L}_T\text{-formula and } \vec{b} \in \mathrm{CT}^M\};$$

note that SPOS_M is different from POS^M (defined in CH.2-8); the extra 'S' stands for 'standard'. We define operators $\mathbf{P}_0$, $\mathbf{P}_1$, $\mathbf{P}_2$ from REG_M to REG_M for PUTB is defined as follows.

(a) $a \in \mathbf{P}_0(X)$ iff $a \in \mathrm{SPOS}_M$ and $a = \mathrm{sb}_{\vec{x}}\big(\ulcorner \phi \urcorner, \vec{b}\big)$ for some positive (standard) $\mathcal{L}_T$-formula $\phi(\vec{x})$ and $\vec{b} \in \mathrm{CT}^M$ such that $X \models \phi\big(\mathrm{val}(\vec{b})\big)$; note that $\mathbf{P}_0(X) \subset \mathrm{SPOS}_M$ for all $X \subset |M|$.

(b) $\mathbf{P}_1$ is defined by

$$\mathbf{P}_1(Z) := \mathbf{P}_0(Z) \cup \big((Z \cap \mathrm{St}^M_{\mathcal{L}_T}) \setminus \mathrm{SPOS}_M\big).$$

(c) $\mathbf{P}_2$ is defined by

$$\mathbf{P}_2(Z) := \mathbf{P}_0(Z) \cup (\mathrm{St}^M_{\mathcal{L}_T} \setminus \mathrm{SPOS}_M).$$

It is easy to see that $\mathbf{P}_0$, $\mathbf{P}_1$ and $\mathbf{P}_2$ are all monotone; for, if $X \subset Y$ and $X \subset \phi(\vec{a})$ then $Y \models \phi(\vec{a})$

for all positive formulae ϕ.

Lemma 4.3.10. $\mathbf{P}_0$, $\mathbf{P}_1$ and $\mathbf{P}_2$ are operators from REG_M to REG_M.

Proof. Let $\phi(x)$ be a T-positive $\mathcal{L}_T$-formula. For each $a \in \mathrm{CT}^M$ and $X \subset |M|$, $X \models \phi\big(\mathrm{val}(a)\big)$ iff $X \models \phi\big(\mathrm{val} \circ \mathrm{nm} \circ \mathrm{val}(c)\big)$. The claim for $\mathbf{P}_0$ follows from this. The other cases trivially follow. $\square$

Lemma 4.3.11. $\mathbf{P}_0$, $\mathbf{P}_1$ and $\mathbf{P}_2$ are all monotone.

Proof. It suffices to show that $\mathbf{P}_0$ is monotone. The monotonicity of $\mathbf{P}_0$ follows from the T-positivity: $X \models \phi$ implies $Y \models \phi$ for all T-positive formula ϕ and $X \subset Y \subset |M|$. $\square$

Lemma 4.3.12. Let $X \in \mathrm{REG}_M$. Then $\mathbf{P}_0(X) \in \mathrm{CONS}_M$ and $\mathbf{P}_2(X) \in \mathrm{COMP}_M$

Proposition 4.3.13. If $X \in \mathrm{CONS}_M$ then $X^{\mathbf{P}_0}_{\mathrm{down}} \in \mathrm{CONS}_M$, and if $X \in \mathrm{COMP}_M$ then $X^{\mathbf{P}_2}_{\mathrm{up}} \in \mathrm{COMP}_M$.

Proof. Since $\mathbf{P}_0(X) \subset \mathrm{POS}_M$ for all X, we have $\mathrm{DOWN}^{\mathbf{P}_0}_M \subset \mathrm{POS}_M$. For simplicity, let us write Y for $\mathrm{DOWN}^{\mathbf{P}_0}(X)$. Then we can show by induction on α that $Y^{\mathbf{P}_0}_\alpha \subset \mathrm{POS}_M \in \mathrm{CONS}_M$ for all ordinals α. For the second claim, we first observe that $\mathrm{UP}^{\mathbf{P}_2}(X) \in \mathrm{COMP}_M$, since $\mathbf{P}_2(X) \in \mathrm{COMP}_M$ for all $X \in \mathrm{REG}_M$ by the last lemma. We write Y for $\mathrm{UP}^{\mathbf{P}_2}(X)$. For the sake of contradiction, suppose $\mathrm{DOWN}^{\mathbf{P}_2}(Y) = X^{\mathbf{P}_2}_{\mathrm{up}} \notin \mathrm{COMP}_M$. Take the least α such that there exists $a \in \mathrm{St}^M_{\mathcal{L}_T}$ with $a, \dot{\neg} a \notin Y^{\alpha+1}_{\mathbf{P}_2}$. Then a or $\dot{\neg} a$ is in $Y^\alpha_{\mathbf{P}_2}$. Assume $a \in Y^\alpha_{\mathbf{P}_2}$. If $a \notin \mathrm{SPOS}_M$ then $a \in Y^{\alpha+1}_{\mathbf{P}_2}$ by definition. Thus, $a \in \mathrm{SPOS}_M$. But then $\dot{\neg} a \in (\mathrm{St}^M_{\mathcal{L}_T} \setminus \mathrm{SPOS}_M)$ and thus $\in Y^{\alpha+1}_{\mathbf{P}_2}$. $\square$

Lemma 4.3.14. $X \models \mathsf{PUTB}^-$ iff $X \in \mathrm{FIX}^{\mathbf{P}_1}_M$. Hence $X \models \mathrm{PUTB}^- + \mathrm{Cons}$ (or Comp) iff $X \in \mathrm{FIX}^{\mathbf{P}_1}_M \cap \mathrm{CONS}_M$ ($\mathrm{FIX}^{\mathbf{P}_1}_M \cap \mathrm{COMP}_M$ resp.).

Proof. The direction from the latter to the former is immediate. Conversely, suppose the former. Then, for every T-positive $\phi(\vec{x})$ and $\vec{a} \in \mathrm{St}^M_{\mathcal{L}_T}$,

$$\mathrm{sb}_{\vec{x}}(\ulcorner \phi \urcorner, \vec{a}) \in X \;\Leftrightarrow\; X \models \phi(\vec{a}^\circ) \;\Leftrightarrow\; \ulcorner \phi(\vec{a}) \urcorner \in \mathbf{P}_1(X).$$

Finally, for each $a \in X \setminus \mathrm{SPOS}_M$, we trivially have $a \in X \Leftrightarrow a \in \mathbf{P}_1(X)$. $\square$

Proposition 4.3.15. $\emptyset_{\mathrm{up}}^{\mathbf{P}_0} = \emptyset_{\mathrm{up}}^{\mathbf{P}_1}$ is the smallest model of PUTB^- and $\mathrm{PUTB}^- + \mathrm{Cons}$, and $(\mathrm{St}^M_{\mathcal{L}_T})^{\mathbf{P}_1}_{\mathrm{down}} = (\mathrm{St}^M_{\mathcal{L}_T})^{\mathbf{P}_2}_{\mathrm{down}}$ is the largest model of PUTB^- and $\mathsf{PUTB}^- + \mathrm{Comp}$.

Proof. We can show the claimed identity by transfinite induction on the constructions of $\emptyset_{\mathrm{up}}^{\mathbf{P}_0}$ and $\emptyset_{\mathrm{up}}^{\mathbf{P}_1}$ (or, $(\mathrm{St}^M_{\mathcal{L}_T})^{\mathbf{P}_1}_{\mathrm{down}}$ and $(\mathrm{St}^M_{\mathcal{L}_T})^{\mathbf{P}_2}_{\mathrm{down}}$, resp.) The minimality and maximality follows from the last lemma and Proposition 4.3.5. $\square$

Next we consider the semantics of schematic reflective closures. The idea is originally due to Feferman [17, pp.22–23]. Let $\mathbf{G}$ be any of $\mathbf{S}$, $\mathbf{F}$, $\mathbf{W}$, $\mathbf{P}_i$ $(i = 0, 1, 2)$. Given $Z \subset |M|$, a new operator $\mathbf{G}^*_Z \colon \mathrm{REG}_M \to \mathrm{REG}_M$ is obtained by (i) adding the following clause to its defining clauses (in Definition 4.3.2 or 4.3.9)

$$M \models a = \dot{P}(b) \wedge \mathrm{CT}^M(b) \wedge b \in Z;$$

and (ii) by replacing the occurrence of '$\mathrm{St}^M_{\mathcal{L}_T}$' in its definition by '$\mathrm{St}^M_{\mathcal{L}_T(P)}$'.

Lemma 4.3.16. **(1)** If $\mathsf{WKF}^* \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{W}^*_Z}_{\mathrm{up}}) \models \phi$ and $(\mathbb{N}, Z, X^{\mathbf{W}^*_Z}_{\mathrm{down}}) \models \phi$ for all $X, Z \subset \mathbb{N}$, where P is interpreted by Z and T is interpreted by $X^{\mathbf{W}^*_Z}_{\mathrm{down}}$. If $\mathsf{WKF}^* + \mathrm{Cons} \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{W}^*_Z}_{\mathrm{down}}) \models \phi$ for all $X \in \mathrm{CONS}_{\mathbb{N}}$. If $\mathsf{WKF}^* + \mathrm{Comp} \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{W}^*_Z}_{\mathrm{up}}) \models \phi$ for all $X \in \mathrm{COMP}_{\mathbb{N}}$.

(2) If $\mathsf{FKF}^* \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{F}^*_Z}_{\mathrm{up}}) \models \phi$ and $(\mathbb{N}, Z, X^{\mathbf{F}^*_Z}_{\mathrm{down}}) \models \phi$ for all $X, Z \subset \mathbb{N}$. If $\mathsf{FKF}^* + \mathrm{Cons} \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{F}^*_Z}_{\mathrm{down}}) \models \phi$ for all $X \in \mathrm{CONS}_{\mathbb{N}}$. If $\mathsf{FKF}^* + \mathrm{Comp} \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{F}^*_Z}_{\mathrm{up}}) \models \phi$ for all $X \in \mathrm{COMP}_{\mathbb{N}}$.

(3) If $\mathsf{KF}^* \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{S}^*_Z}_{\mathrm{up}}) \models \phi$ and $(\mathbb{N}, Z, X^{\mathbf{S}^*_Z}_{\mathrm{down}}) \models \phi$ for all $X, Z \subset \mathbb{N}$. If $\mathsf{KF}^* + \mathrm{Cons} \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{S}^*_Z}_{\mathrm{down}}) \models \phi$ for all $X \in \mathrm{CONS}_{\mathbb{N}}$. If $\mathsf{KF}^* + \mathrm{Comp} \vdash \phi$ then $(\mathbb{N}, Z, X^{\mathbf{S}^*_Z}_{\mathrm{up}}) \models \phi$ for all $X \in \mathrm{COMP}_{\mathbb{N}}$.

Lemma 4.3.17. If $\mathsf{PUTB}^* \vdash \phi$ then $(\mathbb{N}, Z, X_{\mathrm{up}}^{\mathbf{G}^*_Z}) \models \phi$ for all $X \subset \mathbb{N}$, where $\mathbf{G}$ can be either of $\mathbf{P}_0$, $\mathbf{P}_1$ and $\mathbf{P}_2$. If $\mathsf{PUTB}^* + \mathrm{Cons} \vdash \phi$ then $(\mathbb{N}, Z, X_{\mathrm{down}}^{(\mathbf{P}_0)^*_Z}) \models \phi$ for all $X \in \mathrm{CONS}_{\mathbb{N}}$. If $\mathsf{PUTB}^* + \mathrm{Comp} \vdash \phi$ then $(\mathbb{N}, Z, X_{\mathrm{up}}^{(\mathbf{P}_2)^*_Z}) \models \phi$ for all $X \in \mathrm{COMP}_{\mathbb{N}}$.

Lemma 4.3.18. If $\mathsf{PUTB}^* + \mathrm{Cons} \vdash T^{\ulcorner}\sigma^{\urcorner}$, then σ is T-positive. Thus, the same can be said to PUTB^- (+Cons) and PUTB (+Cons).

Proof. This is shown by induction on α that, for any non T-positive $\phi(\vec{x})$, $\emptyset_\alpha^{(\mathbf{P}_0)^*_\emptyset}$ (with respect to the standard model $\mathbb{N}$) contains no $\mathrm{sb}_{\vec{x}}(^{\ulcorner}\phi^{\urcorner}, \vec{a})$ for any $\vec{a} \in \mathbb{N}$. □

Theorem 4.3.19. If $\mathsf{PUTB}^* \vdash T^{\ulcorner}\sigma^{\urcorner}$ then $\mathsf{PUTB}^* \vdash \sigma$. Thus, $\mathcal{I}(\mathsf{PUTB}^*) \subset \mathsf{PUTB}^*$ and PUTB^* + CONEC is identical with PUTB^*. The same applies even to PUTB and PUTB^-.

Proof. It immediately follows from the last lemma, since $\mathsf{PUTB}^* \vdash T^{\ulcorner}\sigma^{\urcorner}$ then σ is T-positive and thus σ is derivable in PUTB^* by the PUTB-schema. □

4.4 Inner Theory Comparison via Semantics

In the present section, we will compare the systems of truth by their inner theories, in other words, by their assertable truths. As we have mentioned in §1, the distinction of the outer and inner theories was first introduced by Reinhardt [70]. Given a system Q over $\mathcal{L}_T$, the outer theory of Q denotes $\{\sigma \in \mathcal{L}_T \mid \mathsf{Q} \vdash \sigma\}$, i.e., the 'theory' in the ordinary sense; on the other hand, the inner theory of Q denotes $\{\sigma \in \mathcal{L}_T \mid \mathsf{Q} \vdash T(^{\ulcorner}\sigma^{\urcorner})\}$, i.e., the set of assertable truths in Q in other words. Let $\mathcal{I}(\mathsf{Q})$ denote the inner theory of a given Q. Then, note that if $\mathcal{I}(\mathsf{Q}_0) \subset \mathcal{I}(\mathsf{Q}_1)$, and if Q_0 and Q_1 prove the T-schema for truth-free sentences, then Q_1 proves as many truth-free theorems as Q_1 does. However, as we shall see, the converse does not hold. Thus, since all the theories we have considered so far prove the T-schema for truth-free sentences, comparison by the inner theories gives properly stronger and finer-grained criterion than the mere derivability of truth-free theorems does.

For the sake of simplicity of the writing, we introduce some notation; for theories Q_0 and Q_1 over $\mathcal{L}_T$, we write

- $\mathsf{Q}_0 \leq_{\mathcal{I}} \mathsf{Q}_1$, when if $\mathsf{Q}_0 \vdash Tt$ then $\mathsf{Q}_1 \vdash Tt$ for all closed term t (thus, if $\mathsf{Q}_0 \leq_{\mathcal{I}} \mathsf{Q}_1$ then $\mathcal{I}(\mathsf{Q}_0) \subset \mathcal{I}(\mathsf{Q}_1)$),
- $\mathsf{Q}_0 \equiv_{\mathcal{I}} \mathsf{Q}_1$, if $\mathsf{Q}_0 \leq_{\mathcal{I}} \mathsf{Q}_1$ and $\mathsf{Q}_1 \leq_{\mathcal{I}} \mathsf{Q}_0$,
- $\mathsf{Q}_0 <_{\mathcal{I}} \mathsf{Q}_1$, if $\mathsf{Q}_0 \leq_{\mathcal{I}} \mathsf{Q}_1$ and $\mathsf{Q}_1 \not\leq_{\mathcal{I}} \mathsf{Q}_0$, and
- $\mathsf{Q}_0 \not>_{\mathcal{I}} \mathsf{Q}_1$, if $\mathsf{Q}_0 \not\leq_{\mathcal{I}} \mathsf{Q}_1$ and $\mathsf{Q}_1 \not\leq_{\mathcal{I}} \mathsf{Q}_0$.

Proposition 4.4.1. **(1)** $\mathsf{WKF}^- <_{\mathcal{I}} \mathsf{FKF}^-$. **(2)** $\mathsf{FKF}^- <_{\mathcal{I}} \mathsf{KF}^-$. **(3)** $\mathsf{KF}^- \equiv_{\mathcal{I}} \mathsf{KF}^- + \mathrm{Cons}$. **(4)** $\mathsf{FKF}^- \equiv_{\mathcal{I}} \mathsf{FKF}^- + \mathrm{Cons}$. **(5)** $\mathsf{WKF}^- \equiv_{\mathcal{I}} \mathsf{WKF}^- + \mathrm{Cons}$. **(6)** $\mathsf{KF}^- <_{\mathcal{I}} \mathsf{KF}^- + \mathrm{Comp}$.

Proof. First, fix an arbitrary model M of PA and let Q^- be one of the above theories and $\mathbf{G}$ be the operator on M for Q^-. The crucial point for the proofs is the fact that if $X \models \mathsf{Q}^-$ then $\emptyset^{\mathbf{G}}_{\mathrm{up}} \subset X$. Thus, (3)-(5) immediately follows from this fact, since $\emptyset^{\mathbf{G}}_{\mathrm{up}} \models \mathrm{Cons}$.

For (1) and (2), it then suffices to show that $\emptyset^{\mathbf{W}}_{\mathrm{up}} \subsetneq \emptyset^{\mathbf{F}}_{\mathrm{up}} \subsetneq \emptyset^{\mathbf{S}}_{\mathrm{up}}$. We can show by transfinite induction that $\emptyset^{\mathbf{W}}_{\alpha} \subset \emptyset^{\mathbf{F}}_{\alpha} \subset \emptyset^{\mathbf{S}}_{\alpha}$ for each α; thus $\emptyset^{\mathbf{W}}_{\mathrm{up}} \subset \emptyset^{\mathbf{F}}_{\mathrm{up}} \subset \emptyset^{\mathbf{S}}_{\mathrm{up}}$. In order to show that these inclusions are proper, consider the following two sentences:

$$\tau_0 :\equiv 0 = 0 \vee \lambda, \text{ where } \lambda \text{ is the liar sentence (defined in Ch.2.6);}$$

$$\tau_1 :\equiv T\forall x\big[\exists y, z(\mathrm{CT}(y) \wedge \mathrm{CT}(z) \wedge x = (y \dot{=} z) \wedge \mathrm{val}(y) = \mathrm{val}(z)) \to Tx\big]$$

Then, we can show by induction on α that $\tau_0 \in \emptyset^{\mathbf{S}}_{\alpha}$ but $\tau_0 \notin \emptyset^{\mathbf{F}}_{\alpha}$, and $\tau_1 \in \emptyset^{\mathbf{F}}_{\alpha}$ but $\tau_1 \notin \emptyset^{\mathbf{W}}_{\alpha}$ for a sufficiently large α; cf. Remark 5 in Ch.2.6. We have obtained the claim.

Finally for (6), since KF^- is a subtheory of $\mathrm{KF}^- + \mathrm{Comp}$, it suffices to show that $\mathsf{KF}^- \not\geq_T \mathsf{KF}^- + \mathrm{Comp}$. This follows from (3) and Prop 2.6.4. □

Proposition 4.4.2. **(1)** $\mathrm{PUTB}^- <_{\mathcal{I}} \mathrm{KF}^-$. **(2)** $\mathrm{PUTB}^- \equiv_{\mathcal{I}} \mathrm{PUTB}^- + \mathrm{Cons}$. **(3)** $\mathrm{PUTB}^- <_{\mathcal{I}} \mathrm{PUTB}^- + \mathrm{Comp}$. **(4)** $\mathrm{PUTB}^- + \mathrm{Comp} \not\sim_{\mathcal{I}} \mathrm{KF}^-$. **(5)** $\mathrm{PUTB}^- \not\sim_{\mathcal{I}} \mathrm{WKF}^-$. **(6)** $\mathrm{PUTB}^- \not\sim_{\mathcal{I}} \mathrm{FKF}^-$. **(7)** $\mathrm{PUTB}^- + \mathrm{Comp} <_{\mathcal{I}} \mathrm{KF}^-$. **(8)** $\mathrm{WKF}^- \not\leq_{\mathcal{I}} \mathrm{PUTB}^- + \mathrm{Comp}$. **(9)** $\mathrm{FKF}^- \not\leq_{\mathcal{I}} \mathrm{PUTB}^- + \mathrm{Comp}$.

Proof. **(1).** It is trivial that $\mathsf{PUTB} \leq_{\mathcal{I}} \mathsf{KF}^-$, since PUTB^- is a subtheory of KF^-. To show that it is a proper subtheory, we note that $\ulcorner \neg T(\ulcorner 0 \neq 0 \urcorner) \urcorner \notin \emptyset_{\mathrm{up}}^{\mathbf{P}_0}$ but it is provable in KF^-.

(2). It is because $\emptyset_{\mathrm{up}}^{\mathbf{P}_0}$ is the least model for both PUTB^- and $\mathrm{PUTB}^- + \mathrm{Cons}$.

(3) & (4). $\mathrm{PUTB}^- + \mathrm{Comp} \vdash \lambda$ but $\mathrm{KF}^- \nvdash \lambda$.

(5) & (6). It suffices to show that $\mathrm{PUTB}^- \not\leq_{\mathcal{I}} \mathrm{FKF}^-$ and $\mathrm{WKF}^- \not\leq_{\mathcal{I}} \mathrm{PUTB}^-$. The former follows from the fact that a sentence $T(\ulcorner 0 = 0 \vee \lambda \urcorner)$ is derivable from PUTB^- but not contained in $\emptyset_{\mathrm{up}}^{\mathbf{F}}$. The latter obtains by observing that the sentence $T\big(\ulcorner \neg T(\ulcorner 0 \neq 0 \urcorner) \urcorner\big)$ is derivable from WKF^- but not contained in $\emptyset_{\mathrm{up}}^{\mathbf{P}_0}$.

(7). Let τ be a sentence $T(\ulcorner 0 = 0 \urcorner) \wedge \neg T(\ulcorner 0 \neq 0 \urcorner)$. We have $\mathsf{KF}^- \vdash T\ulcorner\tau\urcorner$. Hence, it suffices to show that $\mathsf{PUTB}^- + \mathrm{COMP} \nvdash T\ulcorner\tau\urcorner$. Define $X := \mathrm{St}_{\mathcal{L}_T}^M \setminus \{\ulcorner\tau\urcorner\}$. Note that neither τ nor $\neg\tau$ is T-positive. First, we can show by induction on α that $\ulcorner\tau\urcorner \notin X_{\mathbf{P}_1}^{\alpha}$ for all α; thus $\ulcorner\tau\urcorner \notin \mathrm{DOWN}^{\mathbf{P}_1}(X)$. Second, we can similarly show that $\ulcorner\tau\urcorner \notin \mathrm{UP}^{\mathbf{P}_1}\big(\mathrm{DOWN}^{\mathbf{P}_1}(X)\big) = X_{\mathrm{down}}^{\mathbf{P}_1}$. Finally, we will show that $X_{\mathrm{down}}^{\mathbf{P}_1} \in \mathrm{COMP}_M$. For any $a \in \mathrm{St}_{\mathcal{L}_T}^M \setminus \mathrm{SPOS}_M$ and $Y \subset \mathrm{St}_{\mathcal{L}_T}^M$, $a \in Y$ entails $a \in \mathbf{P}_1(Y)$; therefore, we have

$$\mathrm{DOWN}^{\mathbf{P}_1}(Y) \setminus \mathrm{SPOS}_M \;=\; \mathrm{UP}^{\mathbf{P}_1}(Y) \setminus \mathrm{SPOS}_M \;=\; Y \setminus \mathrm{SPOS}_M,$$

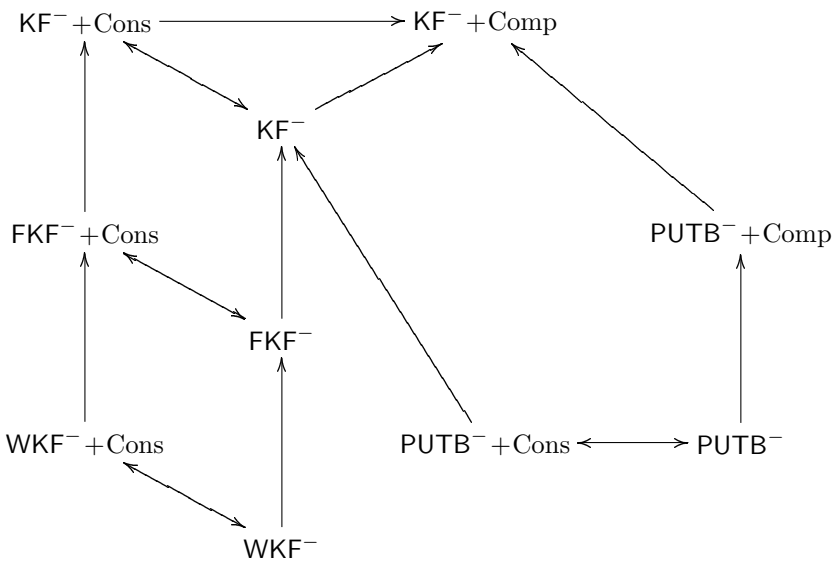
in general. Hence, $\mathrm{St}_{\mathcal{L}_T}^M \setminus (\mathrm{SPOS}_M \cup \{\ulcorner\tau\urcorner\}) \subset X_{\mathrm{down}}^{\mathbf{P}_1}$ Then, since $\neg\tau$ is not T-positive, we have $\ulcorner\neg\tau\urcorner \in X_{\mathrm{down}}^{\mathbf{P}_1}$; thus we obtain $X_{\mathrm{down}}^{\mathbf{P}_1} \in \mathrm{COMP}_M$.

(8) & (9). Immediate from the proof of (7), since $T\ulcorner\tau\urcorner$ is derivable both in WKF^- and FKF^-. □

It follows from the arguments so far that all the systems of truth are distinct except the cases which we have already remarked.

Proposition 4.4.3. The following are all distinct theories: KF^-, FKF^-, WKF^-, $\mathrm{KF}^- + \mathrm{Cons}$, $\mathrm{KF}^- + \mathrm{Comp}$, $\mathrm{FKF}^- + \mathrm{Cons}$, $\mathrm{WKF}^- + \mathrm{Cons}$, PUTB^-, $\mathrm{PUTB}^- + \mathrm{Cons}$, and $\mathrm{PUTB}^- + \mathrm{Comp}$. They all are still distinct even when we add the full induction to them.

We have thus completely correlated all the systems without full-induction in terms of $<_{\mathcal{I}}$, $\leq_{\mathcal{I}}$, $\equiv_{\mathcal{I}}$ and $\not\sim_{\mathcal{I}}$. The results are summarized in the next diagram.



This diagram is read as follows: regarding the diagram as a directed graph,

- if there exists a path from a system Q_0 to a system Q_1, then $\mathsf{Q}_0 \leq_{\mathcal{I}} \mathsf{Q}_1$, and
- if there is no path from a system Q_0 to a system Q_1, then $\mathsf{Q}_0 \not\leq_{\mathcal{I}} \mathsf{Q}_1$.

We can see from this diagram that even when the outer theories are different, the inner theories may coincide; e.g., $\mathcal{O}(\mathsf{KF} + \mathrm{Cons}) \neq \mathcal{O}(\mathsf{KF})$ but $\mathcal{I}(\mathsf{KF} + \mathrm{Cons}) = \mathcal{I}(\mathsf{KF})$.

We have considered systems without full induction (for $\mathcal{L}_T$). When we turn our eyes to systems with it, the situation will become harder to deal with. We cannot make use of the arguments

developed in the present section. For example, even when $M \models \mathsf{PA}$ can be expanded to a model of WKF, it is still open whether M is also expandable to a model of KF, though the converse does hold as was shown in the last chapter. For another example, even when $M \models \mathrm{PA}$ is expandable to a model of KF^-, it is not yet known whether $\emptyset^{\mathbf{S}}_{\mathrm{up}}$ is a model of KF. However, only some partial results on their inner theories can be shown using the so far developed semantical arguments. The other cases will be shown in the subsequent sections by totally different techniques.

Proposition 4.4.4. The following hold:

(1) $\mathsf{KF}^{(*)} <_{\mathcal{I}} \mathsf{KF}^{(*)} + \mathrm{Comp}$, $\mathsf{FKF}^{(*)} <_{\mathcal{I}} \mathsf{FKF}^{(*)} + \mathrm{Comp}$, $\mathsf{WKF}^{(*)} <_{\mathcal{I}} \mathsf{WKF}^{(*)} + \mathrm{Comp}$ and $\mathsf{PUTB}^{(*)} <_{\mathcal{I}} \mathsf{PUTB}^{(*)} + \mathrm{Comp}$

(2) $\mathsf{PUTB}^{(*)} <_{\mathcal{I}} \mathsf{KF}^{(*)}$

(3) $\mathsf{PUTB}^{(*)} \not\sim_{\mathcal{I}} \mathsf{WKF}^{(*)}$, $\mathsf{PUTB}^{(*)} \not\sim_{\mathcal{I}} \mathsf{FKF}^{(*)}$, and $\mathsf{KF}^{(*)} \not\sim_{\mathcal{I}} \mathsf{PUTB}^{(*)} + \mathrm{Comp}$.

Proof. By Theorems 4.3.8, 4.3.14 and 4.3.16, we can construct the same unprovable sentences within (a certain expansion of) the standard model $\mathbb{N}$ even for the cases of systems with full-induction and their schematic reflective closures. □

4.5 Proof theory for finitely iterated self-applicable truths

The proof-theoretic strength of $\mathsf{KF} + \mathrm{Cons}$ was determined by Cantini [5]. In the present section, we will generalize Cantini's technique for other systems WKF and FKF of self-applicable truth and also for systems of iterated self-applicable truth.

As we have seen, the truth predicates of KF, FKF and WKF can be defined as a fixed-point of a certain positive arithmetical operator. Hence, they are all syntactically embeddable in the system $\widehat{\mathsf{ID}}_1$ of fixed-points. With an easy generalization, we can syntactically embed KF_n, FKF_n and WKF_n in the system $\widehat{\mathsf{ID}}_n$ of n iterated fixed-points.

4.5.1 Iterated fixed-points and self-applicable truths

In the present subsection, we introduce the systems $\widehat{\mathsf{ID}}_n$ of n-iterated fixed-points for $n > 0$. The proof-theoretic analyses of $\widehat{\mathsf{ID}}_n$ are well-known and we will make use of them to correlate the inner theories of systems of self-applicable truth.

We recursively define the language $\mathcal{L}_n^{\mathrm{fix}}$ of $\widehat{\mathsf{ID}}_n$ in the following manner. First we set $\mathcal{L}_0^{\mathrm{fix}}$ to be $\mathcal{L}_0$. For each $n \in \mathbb{N}$, let $\mathcal{L}_n^{\mathrm{fix}}(R)$ be $\mathcal{L}_n^{\mathrm{fix}} \cup \{R\}$ where R is a fresh unary predicate symbol. An $\mathcal{L}_n^{\mathrm{fix}}(R)$-formula $\mathcal{A}(R, x)$ is called $(n+1)$-inductive operator form, if $\mathcal{A}$ contains at most x free and R occurs only positively in $\mathcal{A}$. The language $\mathcal{L}_{n+1}^{\mathrm{fix}}$ is defined as $\mathcal{L}_n^{\mathrm{fix}}$ plus *unary* predicates $P_{n+1}^{\mathcal{A}}$ associated to each $(n+1)$-inductive operator form $\mathcal{A}(R, x)$. Then, the system $\widehat{\mathsf{ID}}_{n+1}$ comprises PA, full-induction for $\mathcal{L}_{n+1}^{\mathrm{fix}}$ and the fixed-point axiom schema: for each $(n+1)$-inductive operator form $\mathcal{A}(R, x)$,

$$\forall x[P_{n+1}^{\mathcal{A}}(x) \leftrightarrow \mathcal{A}(P_{n+1}^{\mathcal{A}}, x)].$$

Next, we briefly explain how we can embed systems of n-iterated self-applicable truth in $\widehat{\mathsf{ID}}_n$.

First, for the sake of the embedding of KF_n, we recursively define $\mathcal{A}_n(R,x)$ as follows:

$$
\begin{aligned}
&x\in \mathrm{St}_a \wedge \forall y,z\in \mathrm{CT}[x=y\doteq z\to y^\circ = z^\circ]\wedge \forall y,z\in \mathrm{CT}[x=y\dot{\neq} z\to y^\circ \neq z^\circ]\\
&\wedge \bigwedge_{m<n}\forall y\in \mathrm{CT}[x=\underset{\cdot}{T}_m y\to P_m^{\mathcal{A}_m}(y^\circ)]\ \wedge \bigwedge_{m<n}\forall y\in \mathrm{CT}[x=\dot{\neg}\underset{\cdot}{T}_m y)\to \neg P_m^{\mathcal{A}_m}(y^\circ)]\\
&\wedge \forall y\in \mathrm{CT}[x=\underset{\cdot}{T}_n y)\to y^\circ\in R]\ \wedge\ \forall y\in \mathrm{CT}[x=\dot{\neg}\underset{\cdot}{T}_n y)\to \dot{\neg} y^\circ\in R]\\
&\wedge \forall y\in \mathrm{St}_n[(x=\dot{\neg}\dot{\neg} y)\to y\in R]\\
&\wedge \forall y,z\in \mathrm{St}_n[(x=y\dot{\wedge} z)\to Ry\wedge Rz]\ \wedge\ \forall y,z\in \mathrm{St}_n\big[(x=\dot{\neg}(y\dot{\wedge} z))\to \big(\dot{\neg} y\in R\vee \dot{\neg} z\in R\big)\big]\\
&\wedge \forall y,z\in \mathrm{St}_n\big[(x=y\dot{\to} z)\to \big(R(\dot{\neg} y)\vee Rz\big)\big]\ \wedge\ \forall y,z\in \mathrm{St}_n\big[\big(x=\dot{\neg}(y\dot{\to} z)\big)\to \big(Ry\wedge R(\dot{\neg} z)\big)\big]\\
&\wedge \forall y,z[\mathrm{St}_n(\dot{\forall} y.z)\wedge x=\dot{\forall} y.z\to \forall w(z(w)\in R)]\\
&\wedge \forall y,z[\mathrm{St}_n(\dot{\forall} y.z)\wedge x=\dot{\neg}\dot{\forall} y.z\to \exists w(\dot{\neg} z(w)\in R)].
\end{aligned}
$$

Then, we can straightforwardly show that, for each $n\in\mathbb{N}$,

$$\mathsf{KF}_n\vdash \Psi \quad\Rightarrow\quad \widehat{\mathsf{ID}}_n\vdash \Psi[P_1^{\mathcal{A}_1}/T_1,\ldots,P_n^{\mathcal{A}_n}/T_n]. \tag{4.1}$$

We can assume without loss of generality that this derivation of $\Psi[P^{\mathcal{A}_1}/T_1,\ldots,P^{\mathcal{A}_n}/T_n]$ within $\widehat{\mathsf{ID}}_n$ only uses the vocabulary from $\mathcal{L}_0\cup\{P^{\mathcal{A}_1},\ldots,P^{\mathcal{A}_n}\}$. Let $\widehat{\mathsf{ID}}_n(\mathcal{A})$ be the subsystem of $\widehat{\mathsf{ID}}_n$ obtained by restricting the vocabulary to $\mathcal{L}_n(\mathcal{A})=\mathcal{L}_0\cup\{P^{\mathcal{A}_1},\ldots,P^{\mathcal{A}_n}\}$. Thus, the above (4.1) is strengthened to the following:

$$\mathsf{KF}_n\vdash \Psi \quad\Rightarrow\quad \widehat{\mathsf{ID}}_n(\mathcal{A})\vdash \Psi[P_1^{\mathcal{A}_1}/T_1,\ldots,P_n^{\mathcal{A}_n}/T_n]. \tag{4.2}$$

Similarly, we can define $\mathcal{B}_m$ and $\mathcal{C}_m$ ($m\in\mathbb{N}$) such that, for each $n\in\mathbb{N}$,

$$\mathsf{FKF}_n\vdash \Psi \quad\Rightarrow\quad \widehat{\mathsf{ID}}_n(\mathcal{B})\vdash \Psi[P_1^{\mathcal{B}_1}/T_1,\ldots,P_n^{\mathcal{B}_n}/T_n] \tag{4.3}$$

$$\mathsf{WKF}_n \vdash \Psi \quad \Rightarrow \quad \widehat{\mathsf{ID}}_n(\mathcal{C}) \vdash \Psi[P_1^{\mathcal{C}_1}/T_1, \ldots, P_n^{\mathcal{C}_n}/T_n], \tag{4.4}$$

where $\widehat{\mathsf{ID}}_n(\mathcal{B})$ and $\widehat{\mathsf{ID}}_n(\mathcal{C})$ are defined analogously to $\widehat{\mathsf{ID}}_n(\mathcal{A})$ with restricted vocabulary $\mathcal{B}_1, \ldots, \mathcal{B}_n$ and $\mathcal{C}_1, \ldots, \mathcal{C}_n$ respectively; their languages are denoted by $\mathcal{L}_n(\mathcal{B})$ and $\mathcal{L}_n(\mathcal{C})$ respectively. For saving space, I explicitly give only the definition of $\mathcal{C}_n$:

$$\begin{aligned}
&x \in \mathrm{St}_a \wedge \forall y, z \in \mathrm{CT}[x = y \dot{=} z \to y^\circ = z^\circ] \wedge \forall y, z \in \mathrm{CT}[x = y \dot{\neq} z \to y^\circ \neq z^\circ] \\
&\wedge \bigwedge_{m<n} \forall y \in \mathrm{CT}[x = \dot{T}_m y \to P_m^{\mathcal{A}_m}(y^\circ)] \wedge \bigwedge_{m<n} \forall y \in \mathrm{CT}[x = \dot{\neg} \dot{T}_m y) \to \neg P_m^{\mathcal{A}_m}(y^\circ)] \\
&\wedge \forall y \in \mathrm{CT}[x = \dot{T}_n y) \to y^\circ \in R] \wedge \forall y \in \mathrm{CT}[x = \dot{\neg} \dot{T}_n y) \to \dot{\neg} y^\circ \in R] \\
&\wedge \forall y \in \mathrm{St}_n[x = \dot{\neg}\dot{\neg} y \to y \in R] \wedge \forall y, z \in \mathrm{St}_n[x = y \dot{\wedge} z \to Ry \wedge Rz] \\
&\wedge \forall y, z \in \mathrm{St}_n\big[x = \dot{\neg}(y \dot{\wedge} z) \to \big((\dot{\neg} y \in R \wedge \dot{\neg} z \in R) \vee (\dot{\neg} y \in R \wedge z \in R) \vee (y \in R \wedge \dot{\neg} z \in R)\big)\big] \\
&\wedge \forall y, z \in \mathrm{St}_n\big[(x = y \dot{\to} z) \to \big(R(\dot{\neg} y) \vee Rz\big)\big] \wedge \forall y, z \in \mathrm{St}_n\big[\big(x = \dot{\neg}(y \dot{\to} z)\big) \to \big(Ry \wedge R(\dot{\neg} z)\big)\big] \\
&\wedge \forall y, z[\mathrm{St}_n(\dot{\forall} y.z) \wedge x = \dot{\forall} y.z \to \forall w(z(w) \in R)] \\
&\wedge \forall y, z[\mathrm{St}_n(\dot{\forall} y.z) \wedge x = \dot{\neg}\dot{\forall} y.z \to \exists w(\dot{\neg} z(w) \in R)];
\end{aligned}$$

$\mathcal{B}_n$ can be straightforwardly obtained from $\mathcal{C}_n$ simply by making suitable changes on the clauses for the conditional '$\to$'.

4.5.2 Semi-Formal Systems for $\widehat{\mathsf{ID}}_n$

In order to give proof-theoretic analyses to $\widehat{\mathsf{ID}}_n$, we need to make use of a semi-formal system H_n for it in which infinitary derivations are allowed. By (4.2)–(4.4), we can restrict our proof-theoretic analyses to certain subsystems of $\widehat{\mathsf{ID}}_n$ with restricted vocabulary.

In the present subsection, we will formulate the semi-formal systems $\mathsf{H}_n(\mathcal{A})$ over $\mathcal{L}_n(\mathcal{A})$, $\mathsf{H}_n(\mathcal{B})$ over $\mathcal{L}_n(\mathcal{B})$ and $\mathsf{H}_n(\mathcal{C})$ over $\mathcal{L}_n(\mathcal{C})$ for each $n > 0$. They are formulated in Tait-calculus and thus we modify the language by assuming that the formulae are all given in their negative normal forms: i.e.,

formulae are assumed to be constructed from literals by $\wedge$, $\vee$, $\exists$ and $\forall$. As is usual in ω-arithmetic, we also assume that only closed formulae (i.e., sentences) are allowed throughout this section. For more details of the formulation of ω-arithmetic in Tait Calculus, see [64] or [66].

For each $n \in \mathbb{N}$, we first define the semi-formal system $\mathsf{H}_n(\mathcal{A})$ for $\widehat{\mathsf{ID}}_n(\mathcal{A})$ as follows.

Axioms:

Ax0. Γ, A, for a true $\mathcal{L}_{\mathsf{PA}}$-literal A.

Ax1. $\Gamma, \phi(t), \neg\phi(s)$, for an atomic ϕ of the form $P_m(t)$ and closed terms s and t with $s^{\mathbb{N}} = t^{\mathbb{N}}$.

Logical rules and Cut: the same as the ordinary ω-arithmetic; more explicitly,

$$\frac{\Gamma, A}{\Gamma, A \vee B} \qquad \frac{\Gamma, B}{\Gamma, A \vee B} \qquad \frac{\Gamma, A \quad \Gamma, B}{\Gamma, A \wedge B} \qquad \frac{\Gamma, A(t)}{\Gamma, \exists x A(x)} \qquad \frac{\Gamma, A(\bar{n}) \quad \text{for all numerals } \bar{n}}{\Gamma, \forall x A(x)}$$

Fixed-point rules: for each $1 \leq m \leq n$,

$$\frac{\Gamma, \mathcal{A}(P_m^{\mathcal{A}_m}, s)}{\Gamma, P_m^{\mathcal{A}_m}(s)} \qquad \frac{\Gamma, \neg\mathcal{A}(P_m^{\mathcal{A}_m}, s)}{\Gamma, \neg P_m^{\mathcal{A}_m}(s)}$$

For each derivation $\mathcal{D}$ of $\mathsf{H}_n(\mathcal{A})$, we can assign an ordinal to $\mathcal{D}$ as its length in the usual manner. We also assign its cut-rank to $\mathcal{D}$ as the maximum surface complexity of the critical formulae of cut. Then, we write $\mathsf{H}_n(\mathcal{A}) \vdash^{\delta}_{\rho} \Gamma$, if Γ is obtained by a derivation in $\mathsf{H}_n(\mathcal{A})$ with the length β and cut-rank ρ. We also write $\mathsf{H}_n(\mathcal{A}) \vdash^{\leq\delta}_{<\rho} \Gamma$, when $\mathsf{H}_n \vdash^{\delta'}_{\rho'} \Gamma$ for some $\delta' < \delta$ and $\rho' < \rho$. We write $\mathsf{H}_n \vdash^{\delta}_{\star} \Gamma$ (or $\mathsf{H}_\alpha \vdash^{\leq\delta}_{\star} \Gamma$) when there is a derivation of Γ in H_n with the length δ ($< \delta$, resp.) in which all its cut formulae are of the forms $P^{\mathcal{A}}_\beta t$ and $\neg P^{\mathcal{A}}_\beta$ ($\beta < \alpha$).

Next, we analogously define the semi-formal systems $\mathsf{H}_n(\mathcal{B})$ and $\mathsf{H}_n(\mathcal{C})$ for $\widehat{\mathsf{ID}}_n(\mathcal{B})$ and $\widehat{\mathsf{ID}}_n(\mathcal{C})$;

they are obtained from $\mathsf{H}_n(\mathcal{A})$ by replacing the two Fixed-point rules respectively by

$$\frac{\Gamma, \mathcal{B}(P_m^{\mathcal{B}_m}, s)}{\Gamma, P_m^{\mathcal{B}_m}(s)} \qquad \frac{\Gamma, \neg\mathcal{B}(P_m^{\mathcal{B}_m}, s)}{\Gamma, \neg P_m^{\mathcal{B}_m}(s)}$$

and

$$\frac{\Gamma, \mathcal{C}(P_m^{\mathcal{C}_m}, s)}{\Gamma, P_m^{\mathcal{C}_m}(s)} \qquad \frac{\Gamma, \neg\mathcal{C}(P_m^{\mathcal{C}_m}, s)}{\Gamma, \neg P_m^{\mathcal{C}_m}(s)}\ .$$

The aforementioned standard properties are also parallelly provable in $\widehat{\mathsf{ID}}_n(\mathcal{B})$ and $\widehat{\mathsf{ID}}_n(\mathcal{C})$.

The desired basic properties such as Structural Lemma, Inversion Lemma for $\wedge$ and $\forall$, Exportation Lemma for $\vee$ and Numerical Equivalence (cf. [66]), can be shown in the usual manners for those systems. Since Tautology Lemma and Induction Lemma can also be shown in the usual manner, we standardly obtain the following two lemmata; cf. [4, 43, 66].

Lemma 4.5.1. If $\widehat{\mathsf{ID}}_n(\mathcal{A}) \vdash \phi$, then $\mathsf{H}_n(\mathcal{A}) \vdash_{\rho}^{\omega+\omega} \phi$ for some finite $\rho \in \mathbb{N}$. The parallel statement holds between $\widehat{\mathsf{ID}}_n(\mathcal{B})$ and $\mathsf{H}_n(\mathcal{B})$ and between $\widehat{\mathsf{ID}}_n(\mathcal{C})$ and $\mathsf{H}_n(\mathcal{C})$.

Lemma 4.5.2. If $\mathsf{H}_n(\mathcal{A}) \vdash_{\rho}^{\beta} \Gamma$ then $\mathsf{H}_n(\mathcal{A}) \vdash_{1}^{<\varepsilon(\beta)} \Gamma$, where $\varepsilon(\beta)$ is the least ε-number greater than β. The analogous statements hold for $\mathsf{H}_n(\mathcal{B})$ and $\mathsf{H}_n(\mathcal{C})$.

Now, we define a 'ramified' semi-formal system $\mathsf{AH}_n(\mathcal{A})$. The language $\mathcal{L}_n^{\mathsf{A}}$ of $\mathsf{AH}_n(\mathcal{A})$ consists of the sentences (in Tait-style) of $\mathcal{L}_0 \cup \{P_m \mid 1 \leq m < n\} \cup \{P_n^{\eta} \mid \eta\}$, where η ranges over ordinals up to a sufficiently large ordinal, say, Γ_0; in other words, the predicate P_n of the highest level is further typed by ordinals. We drop the superscripts $\mathcal{A}_m$'s $(m \leq n)$; but this will cause no ambiguity.

For each $n \in \mathbb{N}$, the system $\mathsf{AH}_n(\mathcal{A})$ over $\mathcal{L}_n^{\mathsf{A}}$ is defined as follows.

Axioms:

Ax0. Γ, A, for a true $\mathcal{L}_{\mathsf{PA}}$-literal A.

Ax1. $\Gamma, \phi(t), \neg\phi(s)$, for an atomic ϕ of the form $P_m(t)$ $(m < n)$ and closed terms s and t with $s^{\mathbb{N}} = t^{\mathbb{N}}$.

Ax2. $\Gamma, \neg P_n^0(t)$, for any closed term t.

Logical rules and Cut the same as H_n (i.e., those of ω-arithmetic).

Fixed-point rules for P_m $(m < n)$: the same as H_{n-1}.

Fixed-point rules for P_n^η: for closed terms s and t with $|t| = \alpha$,

$$\frac{\Gamma, \mathcal{A}_m(P_m^\eta, s)}{\Gamma, P_m^{\eta+1}(s)} \qquad \frac{\Gamma, \neg\mathcal{A}_m(P_m^\eta, s)}{\Gamma, \neg P_m^{\eta+1}(s)}$$

Limit rules for P_n^λ: for a limit λ and a closed term s,

$$\frac{\Gamma, P_n^\eta(s) \quad (\text{for some } \eta < \lambda)}{\Gamma, P_n^\lambda(s)} \qquad \frac{\Gamma, \neg P_n^\eta(s) \quad (\text{for all } \eta < \lambda)}{\Gamma, \neg P_n^\lambda(s)}$$

For each $\mathcal{L}_n^{\mathsf{A}}$-formula ϕ, we define its n-rank, $rk_n(\phi)$, as follows: If ϕ is a literal of the form $P_n^\eta t$ or $\neg P_n^\eta t$ then $rk_n(\phi) = \omega \cdot \eta$; if ϕ is a literal of any other form, $rk_n(\phi) = 0$; $rk_n(\phi)$ for a compositional formula ϕ is inductively defined in the standard manner; e.g., $rk_n(\psi_0 \wedge \psi_1) := \max\{rk_n(\psi_0), rk_n(\psi_1)\} + 1$. We write $\mathsf{AH}_n(\mathcal{A}) \vdash_\rho^\delta \Gamma$ (or $\mathsf{AH}_n(\mathcal{A}) \vdash_{<\rho}^{\leq\delta} \Gamma$) when Γ is derived in $\mathsf{AH}_n(\mathcal{A})$ with the length δ ($< \delta$ resp.) and cut-rank (w.r.t. rk_n) ρ ($< \rho$, resp.).

The systems $\mathsf{AH}_n(\mathcal{B})$ and $\mathsf{AH}_n(\mathcal{C})$ are analogously defined: i.e., they are obtained from $\mathsf{AH}_n(\mathcal{A})$ by replacing the Fixed-point rules for P_n^η respectively by

$$\frac{\Gamma, \mathcal{B}_m(P_m^\eta, s)}{\Gamma, P_m^{\eta+1}(s)} \qquad \frac{\Gamma, \neg\mathcal{B}_m(P_m^\eta, s)}{\Gamma, \neg P_m^{\eta+1}(s)}$$

and

$$\frac{\Gamma, \mathcal{C}_m(P_m^\eta, s)}{\Gamma, P_m^{\eta+1}(s)} \qquad \frac{\Gamma, \neg\mathcal{C}_m(P_m^\eta, s)}{\Gamma, \neg P_m^{\eta+1}(s)}\ .$$

Also, $\mathsf{AH}_n(\mathcal{B})\vdash_\rho^\delta \Gamma$ (or $\mathsf{AH}_n(\mathcal{B})\vdash_{<\rho}^{\leq\delta} \Gamma$) and $\mathsf{AH}_n(\mathcal{C})\vdash_\rho^\delta \Gamma$ (or $\mathsf{AH}_n(\mathcal{C})\vdash_{<\rho}^{\leq\delta} \Gamma$) are defined in analogous manners.

4.5.3 Asymmetric Interpretation

Let us fix $n > 0$ throughout the present subsection. In the present subsection, for the sake of simplicity and readability, we will focus on systems with the restricted vocabulary $\mathcal{A}_1, \ldots, \mathcal{A}_n$, i.e., $\widehat{\mathsf{ID}}_n(\mathcal{A})$, $\mathsf{H}_n(\mathcal{A})$ and $\mathsf{AH}_n(\mathcal{A})$ ($n \in \mathbb{N}$). However, all the following arguments can be directly applied to the other two cases. We also suppress '$\mathcal{A}$' from the names of these systems and their predicates.

Let A be an $\mathcal{L}_n^{\mathrm{fix}}$-formula. Then, an $\mathcal{L}_n^\infty$-formula $A[\delta, \gamma]$ is the result of replacing each negative and positive occurrences of T_n respectively by P_n^δ and P_n^γ. We particularly notice that $(\neg A)[\delta, \gamma]$ is identical with $\neg\big(A[\gamma, \delta]\big)$.

Lemma 4.5.3. Let $\delta \leq \gamma$. Then, $\mathsf{AH}_n\vdash_{\omega\cdot\delta}^{\omega\cdot\delta} \neg P_n^\delta s, P_n^\gamma t$.

Proof. Induction on δ. The case in which $\delta = 0$ is trivial from the Ax2. Let us assume $\delta > 0$.

First, suppose $\gamma = \gamma' + 1$ and $\delta = \delta' + 1$. Since $\delta' \leq \gamma'$, we obtain by IH

$$\mathsf{AH}_n\vdash_{\omega\cdot\delta'}^{\omega\cdot\delta'} \neg P_n^{\delta'} r_0, P_n^{\gamma'} r_1,$$

for arbitrary terms r_0 and r_1. Using this, we can show by the argument of the so-called Tautology Lemma (cf. [66]) that

$$\mathsf{AH}_n\vdash_{\omega\cdot\delta'}^{\omega\cdot\delta'+k} \neg\mathcal{A}_n(P_n^{\delta'}, s), \mathcal{A}_n(P_n^{\gamma'}, t),$$

for some finite k depending on $sc(\mathcal{A})$ (i.e., the surface complexity of $\mathcal{A}$). We obtain

$$\vdash \frac{\omega \cdot \delta' + k + 2}{\omega \cdot \delta'} \neg P_n^{\delta} s, P_n^{\gamma} t.$$

Second, suppose $\gamma = \gamma' + 1$ and δ is limit. Take an arbitrary $0 < \delta' < \delta$. Since $\delta' < \gamma'$, it follows from IH that

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta'}{\omega \cdot \delta'} \neg P_n^{\delta'} r_0, P_n^{\gamma'} r_1,$$

for arbitrary terms r_0 and r_1. Then, again by the Tautology Lemma argument, we we have

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta' + k}{\omega \cdot \delta'} \neg \mathcal{A}_n(P_n^{\delta'}, s), \mathcal{A}_n(P_n^{\gamma'}, t),$$

for some finite k. Hence, we obtain

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta' + k + 2}{\omega \cdot \delta'} \neg P_n^{\delta'+1} s, P_n^{\gamma}.$$

Since δ is limit, we have $\delta' + 1 < \delta$ and we already have by IH that

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta'}{\omega \cdot \delta'} \neg P_n^{\delta'} s, P_n^{\delta'+1}.$$

Since $\delta' + 1 < \delta$, we obtain by cut that

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta' + k + 3}{\omega \cdot (\delta' + 1) + 1} \neg P_n^{\delta'} s, P_n^{\gamma}.$$

Since $\delta' < \delta$ is an arbitrary and $\omega \cdot \delta' + k' + 3, \omega \cdot (\delta' + 1) + 1 < \omega \cdot \delta$, we finally obtain by the limit

rule that

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta}{\omega \cdot \delta} \ \neg P_n^{\delta} s, P_n^{\gamma}.$$

Third, suppose γ is limit and $\delta = \delta' + 1$. Similarly to the previous cases, we obtain by IH and the Tautology Lemma argument, we have

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta' + k}{\omega \cdot \delta'} \ \neg \mathcal{A}_n(P_n^{\delta'}, s), \mathcal{A}_n(P_n^{\delta'}, t),$$

for some finite k, and we then have

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta' + k + 2}{\omega \cdot \delta'} \ \neg P_n^{\delta} s, P_n^{\delta}.$$

By the limit rule, we obtain

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta' + k + 3}{\omega \cdot \delta'} \ \neg P_n^{\delta} s, P_n^{\gamma} t.$$

Finally suppose γ and δ are both limit. For each $\delta' < \delta$, by IH we have

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta'}{\omega \cdot \delta'} \ \neg P_n^{\delta'} s, P_n^{\delta'} t.$$

By limit rule we have

$$\mathsf{AH}_n \vdash \frac{\omega \cdot \delta' + 1}{\omega \cdot \delta'} \ \neg P_n^{\delta'} s, P_n^{\gamma} t.$$

Since $\delta' < \gamma$ is arbitrary, we obtain by (the other) limit rule again that

$$\mathsf{AH}_n \vdash^{\omega\cdot\delta}_{\omega\cdot\delta} \neg P^\delta_n s, P^\gamma_n t. \qquad \square$$

The following three lemmata can be shown standardly using the last lemma.

Lemma 4.5.4. Let $\delta_0 \leq \delta_1 \leq \gamma_1 \leq \gamma_0$. Then, $\mathsf{AH}_n \vdash^{\omega\cdot\gamma_1+2\cdot sc(A)}_{\omega\cdot\delta_1} A[\delta_0,\gamma_0], \neg A[\gamma_1,\delta_1]$.

Lemma 4.5.5. Let $\delta_0 \leq \delta_1 \leq \gamma_1 \leq \gamma_0$. Suppose $\vdash^\beta_\rho A[\delta_1,\gamma_1], \Gamma$. Then,

$$\mathsf{AH}_n \vdash^{\max\{\omega\cdot\gamma_1+2\cdot sc(A),\ \beta\}+1}_{\max\{\omega\cdot\gamma_1+sc(A)+1,\ \rho\}} A[\delta_0,\gamma_0], \Gamma.$$

Lemma 4.5.6 (Persitency Lemma)**.** Let $\Gamma = \{\theta_1,\ldots,\theta_k\}$ and let $\vec{\delta}, \vec{\delta}', \vec{\gamma}, \vec{\gamma}'$ be such that $\delta'_i \leq \delta_i \leq \gamma_i \leq \gamma'_i$ for each $1 \leq i \leq k$. Then, if $\vdash^\beta_\rho \Gamma[\vec{\delta},\vec{\gamma}], \Delta$, it holds that

$$\mathsf{AH}_n \vdash^{\max\{\omega\cdot|\vec{\gamma}|+2\cdot|\Gamma|,\ \beta\}+k}_{\max\{\omega\cdot|\vec{\gamma}|+|\Gamma|+1,\ \rho\}} \Gamma[\vec{\delta}',\vec{\gamma}'], \Delta,$$

where $|\vec{\gamma}|$ is $\max\{\gamma_1,\ldots,\gamma_k\}$ and $|\Gamma|$ is $\max\{sc(\theta_1),\ldots,sc(\theta_k)\}$.

The next lemma can also be standardly shown by using the previous four lemmata, but I will give some details since it is the main lemma of the present subsection.

Lemma 4.5.7 (Embedding Lemma)**.** Suppose $\mathsf{H}_n \vdash^\delta_\star \Gamma$. Then for each ordinal η,

$$\mathsf{AH}_n \vdash^{\omega(\eta+\omega^\delta)}_{\omega(\eta+\omega^\delta)} \Gamma[\eta,\eta+\omega^\delta].$$

Proof. The claim is shown by induction on δ.

For the base step, suppose $\mathsf{H}_n \vdash^0_\star \Gamma$ is obtained by an axiom. The case where it is obtained by (Ax0) is trivial. Suppose it is obtained by (Ax1). If the critical formulae are in $\mathcal{L}_m$ $(m < n)$, then

the claim obtains by the same axiom. Otherwise, the critical formulae are $P_n t$ and $\neg P_n s$ for some $t^{\mathbb{N}} = s^{\mathbb{N}}$; then, by the previous Lemma and Structural Lemma, we obtain

$$\mathsf{AH}_n \vdash \frac{\omega \cdot (\eta + \omega^0)}{\omega \cdot (\eta + \omega^0)}\ \Gamma[\eta, \eta + \omega^0].$$

In the rest of proof, we assume that $\delta > 0$. We only illustrate some typical cases.

First, suppose the last inference is by the positive fixed-point rule. Then, the premise is of the form:

$$\mathsf{H}_n \vdash \frac{\delta'}{\star}\ \mathcal{A}_n(P_n, s), \Gamma;$$

we thus obtain by IH that

$$\mathsf{AH}_n \vdash \frac{\omega(\eta + \omega^{\delta'})}{\omega(\eta + \omega^{\delta'})}\ \Gamma[\eta, \eta + \omega^{\delta'}], \mathcal{A}_n(P_n^{\eta + \omega^{\delta'}}, s).$$

By Persistency Lemma,

$$\mathsf{AH}_n \vdash \frac{\omega(\eta + \omega^{\delta'}) + k}{\omega(\eta + \omega^{\delta'}) + k}\ \Gamma[\eta, \eta + \omega^{\delta}], \mathcal{A}_n(P_n^{\omega^{\delta'}}, s),$$

for some finite k. Then, by fixed-point rule, we have

$$\mathsf{AH}_n \vdash \frac{\omega(\eta + \omega^{\delta'}) + k + 1}{\omega(\eta + \omega^{\delta'}) + k}\ \Gamma[\eta, \eta + \omega^{\delta}], P_n^{\eta + \omega^{\delta'} + 1}(s).$$

By Persistency Lemma again, we finally obtain that, for some finitek',

$$\vdash \frac{\omega \cdot (\eta + \omega^{\delta'} + 1) + k'}{\omega \cdot (\eta + \omega^{\delta'} + 1) + k'}\ \Gamma\ [\eta, \eta + \omega^{\delta}], P_\beta^{\eta + \omega^{\delta}}(s).$$

Next, suppose the last inference is the negative fixed-point rule: that is, the premise is of the form

$$\mathsf{H}_n\vdash\frac{\delta'}{\star}\ \neg\mathcal{A}_n(P_n,s),\Gamma,$$

for some $\delta'<\delta$. We obtain by IH that

$$\mathsf{AH}_n\vdash\frac{\omega(\eta+\omega^{\delta'})}{\omega(\eta+\omega^{\delta'})}\ \Gamma[\eta,\eta+\omega^{\delta'}],\neg\mathcal{A}_n(P_n^\eta,s).$$

Then, by Persistency Lemma, we have

$$\mathsf{AH}_n\vdash\frac{\omega(\eta+\omega^{\delta'})+k}{\omega(\eta+\omega^{\delta'})+k}\ \Gamma[\eta,\eta+\omega^{\delta}],\neg\mathcal{A}(P_n^\eta,s)$$

for some finite k. By fixed-point rule, we have

$$\mathsf{AH}_n\vdash\frac{\omega(\eta+\omega^{\delta'})+k+1}{\omega(\eta+\omega^{\delta'})+k}\ \Gamma[\eta,\eta+\omega^{\delta}],\neg P_\beta^{\omega^\eta+1}(s).$$

By Persistency Lemma again, we finally obtain

$$\mathsf{AH}_n\vdash\frac{\omega(\eta+\omega^{\delta})+k+2}{\omega(\eta+\omega^{\delta})+k}\ \Gamma[\eta,\eta+\omega^{\delta}],\neg P_\beta^{\eta}(s).$$

As the final and crucial example, suppose the last inference is cut. Thus the premises are of the forms

$$\mathsf{H}_n\vdash\frac{\delta_0}{\star}\ \Gamma,P_nt\qquad\mathsf{H}_n\vdash\frac{\delta_1}{\star}\ \Gamma,\neg P_nt,$$

for some $\delta_0, \delta_1 < \delta$. By IH, we have

$$\mathsf{AH}_n \vdash^{\omega(\eta+\omega^{\delta_0})}_{\omega(\eta+\omega^{\delta_0})} \Gamma[\eta, \eta+\omega^{\delta_0}], P_n^{\eta+\omega^{\delta_0}}(t)$$
$$\mathsf{AH}_n \vdash^{\omega(\eta+\omega^{\delta_0}+\omega^{\delta_1})}_{\omega(\eta+\omega^{\delta_0}+\omega^{\delta_1})} \Gamma[\eta+\omega^{\delta_0}, \eta+\omega^{\delta_0}+\omega^{\delta_1}], \neg P_n^{\eta+\omega^{\delta_0}}(t).$$

Since $\omega^{\delta_0} + \omega^{\delta_1} < \omega^{\delta}$ and thus by Persistency Lemma,

$$\mathsf{AH}_n \vdash^{\omega(\eta+\omega^{\delta'})+k_0}_{\omega(\eta+\omega^{\delta_0})+k_0} P_n^{\eta+\omega^{\delta_0}}(t), \Gamma[\eta, \eta+\omega^{\delta}]$$
$$\mathsf{AH}_n \vdash^{\omega(\eta+\omega^{\delta_0}+\omega^{\delta_1})+k_1}_{\omega(\eta+\omega^{\delta_0}+\omega^{\delta_1})+k_1} \neg P_n^{\eta+\omega^{\delta_0}}(t), \Gamma[\eta, \eta+\omega^{\delta}],$$

for some finite k_0 and k_1. We finally obtain by applying cut that

$$\mathsf{AH}_n \vdash^{\omega(\eta+\omega^{\delta})}_{\omega(\eta+\omega^{\delta})} \Gamma[\eta, \eta+\omega^{\delta}]$$

The other cases can be shown in similar ways and we complete the proof. □

The next two lemmata are all standardly shown; cf. [66].

Lemma 4.5.8 (Reduction Lemma)**.** Suppose $\vdash^{\beta}_{1+\rho} \phi, \Gamma$ and $\vdash^{\gamma}_{1+\rho} \neg\phi, \Gamma$. Then $\vdash^{\beta\#\gamma}_{1+\rho} \Gamma$.

Lemma 4.5.9 (Predicative Cut Elimination)**.** If $\mathsf{AH}_n \vdash^{\delta}_{1+\rho+\omega^{\beta}} \Gamma$, then $\mathsf{AH}_n \vdash^{\varphi_\beta(\delta)}_{1+\rho} \Gamma$.

The next immediately follows from this last lemma.

Lemma 4.5.10. If $\widehat{\mathsf{ID}}_n \vdash \phi$ then $\mathsf{AH}_n \vdash^{<\varphi_{\varepsilon_0}0}_{1} \phi[0, \xi]$ for some $\xi < \varepsilon_0$.

Lemma 4.5.11. If $\mathsf{AH}_n \vdash^{\delta}_{\rho} P_n^0, \Gamma$, then $\mathsf{AH}_n \vdash^{\delta}_{\rho} \Gamma$.

Proof. This is shown by induction on δ; use the fact that an atomic of the form P_n^0 cannot be critical in any derivation in AH_n. □

Lemma 4.5.12 (Partial Cut Elimination)**.** If $\mathsf{H}_n|\!\frac{\delta}{\rho}\Gamma$ (for any $\rho < \omega$), then $\mathsf{H}_n|\!\frac{\varepsilon(\delta)}{\star}$.

Lemma 4.5.13. Let $n > 1$. If $\widehat{\mathsf{ID}}_n \vdash \phi$ for $\phi \in \mathcal{L}_m^{\mathrm{fix}}(\mathcal{A})$ then $\mathsf{H}_{n-1}|\!\frac{<\varphi_{\varepsilon_0}0}{\star}\ \phi$, for some $\xi < \varepsilon_0$.

Proof. Suppose $\widehat{\mathsf{ID}}_n \vdash \phi$. By Lemma 4.5.10, we have $\mathsf{AH}_n|\!\frac{\delta}{1}\ \phi$ for some $\xi < \varepsilon_0$ and $\delta < \varphi_{\varepsilon_0}0$. By Lemma 4.5.11, we can assume without loss of generality that this derivation contains cuts applied only to $\mathcal{L}_{n-1}^{\mathrm{fix}}$-literals. Hence, we have $\mathsf{H}_n|\!\frac{\delta}{1}\ \phi$. Finally, it follows from the last lemma that $\mathsf{H}_n|\!\frac{\varepsilon(\delta)}{\star}\phi$ and $\varepsilon(\delta) < \varphi_{\varepsilon_0}0$. □

Lemma 4.5.14. Let us recursively define $\alpha_0 = \varepsilon_0$ and $\alpha_{k+1} = \varphi_{\alpha_k}0$. Suppose $\widehat{\mathsf{ID}}_n \vdash \phi$ for an $\mathcal{L}_m^{\mathrm{fix}}$-formula ϕ $(m \leq n)$. Then $\mathsf{AH}_m|\!\frac{<\alpha_{n-m+1}}{1}\ \phi[0,\xi]$ for some $\xi < \alpha_{n-m}$.

4.6 Inner theory comparison via semi-formal systems

In the present section, we will present how the inner theories of systems of iterative compositional truth with full-induction are compared.

4.6.1 Systems with and without Cons

The main theorem of the present subsection is:

Theorem 4.6.1. (1) If $\mathsf{KF} + \mathrm{Cons} \vdash Tt$ then $\mathsf{KF} \vdash Tt$. The parallel statements holds between $\mathsf{FKF} + \mathrm{Cons}$ and FKF and between $\mathsf{WKF} + \mathrm{Cons}$ and WKF.

(2) If $(\mathsf{KF} + \mathrm{Cons})^* \vdash Tt$ then $\mathsf{KF}^* \vdash Tt$. The parallel statements holds between $(\mathsf{FKF} + \mathrm{Cons})^*$ and FKF^* and between $(\mathsf{WKF} + \mathrm{Cons})^*$ and WKF^*.

We will show this theorem in the rest of the present subsection.

For the sake of readability, let us simply write Cons_n for $\mathrm{Cons}_n[P_1^{\mathcal{A}_1}/T_1, \ldots, P_n^{\mathcal{A}_n}/T_n]$ in what follows. Again, we will focus on KF_n but the following arguments parallelly apply to WKF_n and FKF_n as well.

First, suppose $\mathsf{KF}_n + \mathrm{Cons} \vdash Tt$. By (4.2), we have $\widehat{\mathsf{ID}}_1 \vdash \mathrm{Cons} \to P_1^{\mathcal{A}_1} t$. Then, it follows from Lemma 4.5.10 that $\mathsf{AH}_1(\mathcal{A}) \vdash_1^{\leq \alpha_1} \neg\mathrm{Cons}[0,\eta], P_1^\eta t$ for some $\eta < \varepsilon_0$; recall that $\alpha_1 := \varphi_{\varepsilon_0} 0$.

Second, for the reflective closure, if $\mathsf{KF}^* + \mathrm{Cons} \vdash Tt$, we have $\mathsf{KF}_n + \mathrm{Cons}_n \vdash Tt$ for some n; then we similarly obtain $\mathsf{AH}_n(\mathcal{A}) \vdash_1^{\leq \alpha_1} \neg\mathrm{Cons}_n, P_1 t$: that is, we have

$$\mathsf{KF}^* + \mathrm{Cons} \vdash Tt \quad \Rightarrow \quad \mathsf{KF}_n \vdash \mathrm{Cons}_n \to Tt \quad \Rightarrow \quad \mathsf{AH}_n(\mathcal{A}) \vdash_1^{< \alpha_1} \neg\mathrm{Cons}_n, P_1 t.$$

The next is the main lemma of the present subsection.

Lemma 4.6.2. Let $\rho > 0$. For arbitrary η_0 and η_1, Suppose that (a) $\mathsf{AH}_n(\mathcal{A}) \vdash_\rho^{\delta_0} P_n^{\eta_0} s, \Gamma$ and (b) $\mathsf{AH}_n(\mathcal{A}) \vdash_\rho^{\delta_1} P_n^{\eta_1} \dot{\neg} s, \Gamma$. Then, we have $\mathsf{AH}_n(\mathcal{A}) \vdash_\rho^{\delta_0 \# \delta_1} \Gamma$, where $\delta_0 \# \delta_1$ is the natural sum of δ_0 and δ_1.

Proof. This lemma informally corresponds to the fact that the minimal Kripkean fixed-point is consistent; in $\mathsf{AH}_n(\mathcal{A})$, each P_n^η corresponds to the η-th stage of the Kripkean construction (of the n-iterated truth) from the empty set, since we start with $\neg P_n^0 t$ for all t. In what follows, we suppress '$\mathcal{A}$'.

We show the claim by induction on $\delta_0 \# \delta_1$. The base case is easy, since an atomic of the form $P_n^\eta t$ cannot be critical in any axiom. For the induction step, the cases where either $P_n^{\eta_0} s$ or $P_n^{\eta_1} \dot{\neg} s$ is not critical in the last inference follow from IH. Let us assume $P_n^{\eta_0} s$ and $P_n^{\eta_1} \dot{\neg} s$ are both critical in the last inferences. When either of the last inferences is by the limit rule, the premises contain $\mathsf{AH}_n \vdash_\rho^{\delta_2} P_n^{\eta_2}(\dot{\neg}) s, \Gamma$ for $\delta_2 < \delta_0$ (or $< \delta_1$) and $\eta_2 < \eta_0$ (or $< \eta_1$); then, the claim follows from IH. Hence, we focus on the essential case in which the last inferences are both the fixed-point rule. Then, we can assume that the premises are

$$\text{(c) } \mathsf{AH}_n \vdash_\rho^{\delta_0'} \Gamma, \mathcal{A}_n(P_\alpha^{\eta_0 - 1}, s) \quad \text{and} \quad \text{(d) } \mathsf{AH}_n \vdash_\rho^{\delta_1'} \Gamma, \mathcal{A}_n(P_\alpha^{\eta_1 - 1}, \dot{\neg} s),$$

for some $\delta_0' < \delta_0$ and $\delta_1' < \delta_1$. By the form of $\mathcal{A}_n$ (and inversion), we can assume that $s \in \mathrm{St}_n$; let

$s = \ulcorner\sigma\urcorner$. The claim is shown by cases according to the form of σ.

Suppose no negation '$\neg$' is attached to σ. We illustrate two cases. First, assume σ is $\phi \wedge \psi$. By $\wedge$-inversion, $\forall$-inversion and $\vee$-exportation, it follows from (c) that

$$\mathsf{AH}_n \vdash^{\delta_0'}_{\rho} \Gamma,\ \neg\mathrm{St}_n(\ulcorner\phi\urcorner),\ \neg\mathrm{St}_n(\ulcorner\psi\urcorner),\ s \neq \ulcorner\phi \wedge \psi\urcorner,\ P_n^{\eta_0 - 1}(\ulcorner\phi\urcorner) \wedge P_n^{\eta_0 - 1}(\ulcorner\psi\urcorner);$$

check with the form of $\mathcal{A}_n$. Since $\neg\mathrm{St}_n(\ulcorner\phi\urcorner), \neg\mathrm{St}_n(\ulcorner\psi\urcorner), s \neq \ulcorner\phi \wedge \psi\urcorner$ are all false $\mathcal{L}_{\mathsf{PA}}$-literals, we have $\mathsf{AH}_n \vdash^{\delta_0'}_{\rho} \Gamma, P_n^{\eta_0 - 1}(\ulcorner\phi\urcorner) \wedge P_n^{\eta_0 - 1}(\ulcorner\psi\urcorner)$. Then, by $\wedge$-inversion,

$$\mathsf{AH}_n \vdash^{\delta_0'}_{\rho} \Gamma, P_n^{\eta_0 - 1}(\ulcorner\phi\urcorner) \quad \text{and} \quad \mathsf{AH}_n \vdash^{\delta_0'}_{\rho} \Gamma, P_n^{\eta_0 - 1}(\ulcorner\psi\urcorner). \tag{4.5}$$

It also follows in the same way from (d) that

$$\mathsf{AH}_n \vdash^{\delta_1'}_{\rho} \Gamma, P_n^{\eta_1 - 1}(\ulcorner\neg\phi\urcorner) \vee P_n^{\eta_1 - 1}(\ulcorner\neg\psi\urcorner). \tag{4.6}$$

Now, we will show by side induction on δ_1' that, under the (main) induction hypothesis up to $< \delta_0 \# \delta_1$, (4.5) and (4.6) (for $\delta_0' < \delta_1$ and $\delta_1' < \delta_1$) implies that $\vdash^{\delta_0' \# \delta_1'}_{\rho} \Gamma$. The base step in which (4.6) is obtained by axiom is trivial. For the induction steps, the essential case is that $P_n^{\eta_1 - 1}(\ulcorner\neg\phi\urcorner) \vee P_n^{\eta_1 - 1}(\ulcorner\neg\psi\urcorner)$ is critical in the last inference; otherwise the claim follows from SIH. In this case, the premise is

$$\mathsf{AH}_n \vdash^{\delta_1''}_{\rho} \Gamma, P_n^{\eta_1 - 1 \ulcorner\neg\phi\urcorner} \quad \text{or} \quad \mathsf{AH}_n \vdash^{\delta_1''}_{\rho} \Gamma, P_n^{\eta_1 - 1 \ulcorner\neg\psi\urcorner},$$

where $\delta_1'' < \delta_1'$. Then our claim follows from IH and (4.5).

Second, assume $s = \ulcorner\sigma\urcorner = \dot{T}_m r$ for $b, r \in \mathrm{CT}$ with $m < n$. By $\wedge$-inversion, $\forall$-inversion and

$\vee$-exportation, it follows from (c) that

$$\mathsf{AH}_n \vdash^{\delta_0'}_{\rho} \Gamma,\ \neg\mathrm{CT}(b),\ \neg\mathrm{CT}(r),\ s \neq \dot{T}_m r,\ P_m(r^\circ).$$

Thus we obtain $\mathsf{AH}_n \vdash^{\delta_0'}_{\rho} \Gamma, P_m(r^\circ)$, since $\neg\mathrm{CT}(b)$, $\neg\mathrm{CT}(r)$ and $s \neq \dot{T}_b r$ are all false $\mathcal{L}_0$-literals. It also follows in the same way from (d) that $\mathsf{AH}_n \vdash^{\delta_1'}_{\rho} \Gamma, \neg P_m(r^\circ)$. Finally, since $\rho > 0$, we obtain by cut that $\mathsf{AH}_n \vdash^{\delta_0 \# \delta_1}_{\rho} \Gamma$. The cases where σ is of another form can be shown similarly; but note that the case where $s = \dot{T}_n r$ needs neither the side-induction step nor cut with rank 0 and the claim directly follows from IH.

Next, assume σ is of the form $\neg\sigma'$; then $\dot{\neg} s = \ulcorner \neg\neg\sigma' \urcorner$. In the same manner as above, we obtain from (d) that $\mathsf{AH}_n \vdash^{\delta_1'}_{\rho} P_n^{\eta_1 - 1}(\ulcorner \sigma \urcorner), \Gamma$. By IH, the claim $\mathsf{AH}_n \vdash^{\delta_0 \# \delta_1}_{\rho} \Gamma$ follows from this and (c); the proof is completed. $\square$

Lemma 4.6.3. (1) Let $\rho > 0$. If $\mathsf{AH}_n(\mathcal{A}) \vdash^{\delta}_{\rho} P_n^{\eta_0} s \wedge P_n^{\eta_1} \dot{\neg} s, \Gamma$, then $\mathsf{AH}_n(\mathcal{A}) \vdash^{\delta \# \delta}_{\rho} \Gamma$.

(2) Let $\rho > 0$. If $\mathsf{AH}_n(\mathcal{A}) \vdash^{\delta}_{\rho} \neg\mathrm{Cons}_n, \Gamma$, then $\mathsf{AH}_n(\mathcal{A}) \vdash^{\delta \# \delta}_{\rho} \Gamma$.

Now we have so far obtained the following:

$$\mathsf{KF} + \mathrm{Cons} \vdash Tt \quad \Rightarrow \quad \mathsf{AH}_1(\mathcal{A}) \vdash^{< \alpha_1}_{1} \neg\mathrm{Cons}[0, \eta], P_1^\eta t \quad \Rightarrow \quad \mathsf{AH}_1(\mathcal{A}) \vdash^{< \alpha_1}_{1} P_1^\eta t,$$

for some $\eta < \varepsilon_0$. For the reflective closures, we now have:

$$\mathsf{KF}^* + \mathrm{Cons} \vdash Tt \ \Rightarrow\ \mathsf{AH}_n(\mathcal{A}) \vdash^{< \alpha_1}_{1} P_1 t \ \Rightarrow\ \mathsf{H}_{n-1}(\mathcal{A}) \vdash^{< \alpha_1}_{\star} P_1 t \ \Rightarrow\ \mathsf{AH}_{n-1}(\mathcal{A}) \vdash^{< \alpha_2}_{1} P_1 t.$$

By repeating this procedure n-times, we obtain:

$$\mathsf{KF}^* + \mathrm{Cons} \vdash Tt \quad \Rightarrow \quad \mathsf{AH}_1 \vdash^{< \alpha_{n+1}}_{1} P_1^\eta t,$$

for some $\eta < \alpha_n$ $(< \Gamma_0)$.

Hence, it suffices for our purpose to show that

(I) KF proves that $\mathsf{AH}_{n-1}\vdash^{\delta}_{1} P_1^{\eta}t$ implies Tt for $\delta < \alpha_1$ and $\eta < \alpha_0 (= \varepsilon_0)$.

(II) KF^* proves that $\mathsf{AH}_1\vdash^{\delta}_{1} P_1^{\eta}t$ implies Tt for $\delta < \alpha_{n+1}$ and $\eta < \alpha_n$;

these are shown by suitably modeling the infinitary derivations in question within KF and KF^*; Recall that given a derivation of $\mathsf{KF}+\mathrm{Cons} \vdash Tt$ or $(\mathsf{KF}+\mathrm{Cons})^* \vdash Tt$, we can primitive recursively compute the δ and η below the above bounds.

The next is well-known.

Theorem 4.6.4. $\mathsf{KF} \vdash \forall a \in tc\mathrm{TI}\big(\beta; Ta(x)\big)$ for any $\beta < \alpha_1$ $(= \varphi_{\varepsilon_0}0)$, and $\mathsf{KF}^* \vdash \forall a \in tc\mathrm{TI}\big(\beta; Ta(x)\big)$ for any $\beta < \Gamma_0$.

By diagonalization, we define a predicate $H_\gamma(x)$ (with the parameter γ and x) represents the

truth of γ-th stage in its Kripke construction. For $\gamma = 0$, set $H_0 = \emptyset$; then,

$$
\begin{aligned}
H_{\gamma+1}(a) :\Leftrightarrow & \exists x, y \in \mathrm{CT}[a = x \dot{=} y \wedge x^\circ = y^\circ] \vee \exists x, y \in \mathrm{CT}[a = x \dot{\neq} y \wedge x^\circ \neq y^\circ] \\
& \vee \exists b, c \in \mathrm{St}_{\mathcal{L}_T}\big[a = b \dot{\wedge} c \wedge \big(T^\ulcorner H_{\dot{\gamma}}(\dot{b})^\urcorner \wedge T^\ulcorner H_{\dot{\gamma}}(\dot{c})^\urcorner\big)\big] \\
& \vee \exists b, c \in \mathrm{St}_{\mathcal{L}_T}\big[a = \dot{\neg}(b \dot{\wedge} c) \wedge \big(T^\ulcorner H_{\dot{\gamma}}(\dot{\neg}\dot{b})^\urcorner \vee T^\ulcorner H_{\dot{\gamma}}(\dot{\neg}\dot{c})^\urcorner\big)\big] \\
& \vee \exists b, c \in \mathrm{St}_{\mathcal{L}_T}\big[a = b \dot{\rightarrow} c \wedge \big(T^\ulcorner H_{\dot{\gamma}}(\dot{\neg}\dot{b})^\urcorner \vee T^\ulcorner H_{\dot{\gamma}}(\dot{c})^\urcorner\big)\big] \\
& \vee \exists b, c \in \mathrm{St}_{\mathcal{L}_T}\big[a = \dot{\neg}(b \dot{\rightarrow} c) \wedge \big(T^\ulcorner H_{\dot{\gamma}}(\dot{b})^\urcorner \wedge T^\ulcorner H_{\dot{\gamma}}(\dot{\neg}\dot{c})^\urcorner\big)\big] \\
& \vee \exists b \in \mathrm{St}_{\mathcal{L}_T}[a = \dot{\neg}\dot{\neg} b \wedge T^\ulcorner H_{\dot{\gamma}}(\dot{b})^\urcorner] \\
& \vee \exists x \in \mathrm{CT}\big[a = \dot{T} x \wedge T^\ulcorner H_{\dot{\gamma}}(\dot{x^\circ})^\urcorner\big] \\
& \vee \exists x \in \mathrm{CT}\big[a = \dot{\neg}\dot{T} x \wedge T^\ulcorner H_{\dot{\gamma}}(\dot{\neg}\dot{x^\circ})^\urcorner\big] \\
& \vee \exists b \exists c[a \in \mathrm{St}_{\mathcal{L}_T} \wedge a = \dot{\forall} c.b \wedge \forall y T^\ulcorner H_{\dot{\gamma}}(\dot{b}(\dot{y})^\urcorner] \\
& \vee \exists b \exists c[a \in \mathrm{St}_{\mathcal{L}_T} \wedge a = \dot{\neg}\dot{\forall} c.b \wedge \exists y T^\ulcorner H_{\dot{\gamma}}(\dot{\neg}\dot{b}(\dot{y})^\urcorner] \\
& \vee a \in H_\gamma;
\end{aligned}
$$

for a limit λ, $H_\lambda(a)$ iff $a \in \bigcup_{\xi<\lambda} H_\xi$; more precisely, $(\exists n) T^\ulcorner H_{\dot{\lambda}[\dot{n}]}(\dot{a})^\urcorner$, where $\lambda[n]$ is the fundamental sequence of λ.

Lemma 4.6.5. We can show by transfinite induction up to ε_0 for $\mathcal{L}_T$ that, for each given $\beta < \varepsilon_0$,

$\mathsf{KF} \vdash \forall \gamma < \bar{\beta} \forall x \big[(T^\ulcorner H_{\dot{\gamma}}(\dot{x})^\urcorner \to Tx) \wedge (H_\gamma(x) \to Tx)\big]$

$\mathsf{KF} \vdash \forall \gamma < \bar{\beta} \forall x \big[(T^\ulcorner H_{\dot{\gamma}}(\dot{x})^\urcorner \to D^+x) \wedge (H_\gamma(x) \to D^+x)\big]$

$\mathsf{KF} \vdash \forall \gamma < \bar{\beta} \forall x \big[\ulcorner H_\gamma(\dot{x})^\urcorner \in D^+ \wedge \ulcorner T^\ulcorner H_{\dot{\gamma}}(\dot{x})^\urcorner\urcorner \in D^+\big]$: thus, $\ulcorner H_{\dot{\gamma}}(x)^\urcorner \in tc$ for all $\gamma < \bar{\beta}$

$\mathsf{KF} \vdash \forall \gamma < \bar{\beta} \forall x [T^\ulcorner H_{\dot{\gamma}}(\dot{x})^\urcorner \leftrightarrow H_\gamma(x)]$.

Given $\alpha < \varepsilon_0$, we define a translation $\mathcal{H}_\alpha$ from $\mathcal{L}_1^{\mathsf{A}}$ into $\mathcal{L}_T$ by $T_\beta x \mapsto H_\beta x$ for $\beta < \alpha$ and

$T_\gamma x \to \ulcorner 0 = 0 \urcorner$ for $\gamma \not< \alpha$; all the other vocabulary and logical symbols are preserved. Let h_α be a (primitive recursive) representation of $\mathcal{H}_\alpha$ (in, say, PA). For a technical reason, we stipulate that $h_\alpha(x) = \ulcorner 0 = 0 \urcorner$ for all $x \notin \mathcal{L}_1^{\mathsf{A}}$.

Corollary 4.6.6. Let $\alpha < \varepsilon_0$. The following holds within KF: for any $\mathcal{L}_1^{\mathsf{A}}$-sentence σ, $\mathcal{H}_\alpha(\sigma)$ is in D^+. More precisely, we have

$$\mathsf{KF} \vdash (\forall x \in \mathrm{St}_{\mathcal{L}_1^{\mathsf{A}}})\big(h_{\bar{\alpha}}(\ulcorner \sigma \urcorner) \in D^+\big);$$

therefore, we have $h_{\bar{\alpha}}(x) \in D^+$ for all x and thus $\ulcorner T h_{\bar{\alpha}}(x) \urcorner \in tc$.

Proof. This is shown by induction on $rk_1(\sigma)$ using the last lemma. □

For $\alpha < \varepsilon_0$, let $\mathcal{L}_1^\alpha := \mathcal{L}_0 \cup \{P_1^\eta \mid \eta < \alpha\}$ $(\subset \mathcal{L}_1^{\mathsf{A}})$. Suppose $\mathsf{AH}_1 \vdash_1^\delta \Gamma$ for $\Gamma \subset \mathcal{L}_1^\alpha$. Then, we can assume by Lemma 4.5.11 that all sequents appearing in its derivation are from $\mathcal{L}_1^\alpha$. Let $\mathsf{ah}(\delta, C)$ is a standard representation (in PA) of $\mathsf{AH}_1 \vdash_1^\delta \Gamma$, where C is the code of the sequent Γ; let $C \subset \mathrm{St}_{\mathcal{L}_1^{\mathsf{A}}}$ express that C is a code of a sequent from $\mathcal{L}_1^{\mathsf{A}}$. We set $\Phi_\alpha(\delta)$ to be the following $\mathcal{L}_T$-formula:

$$\forall C \subset \mathrm{St}_{\mathcal{L}_1^{\mathsf{A}}}\big[\mathsf{ah}_1(\delta, C) \to T\big(h_{\bar{\alpha}}(\dot{\bigwedge} C)\big)\big]$$

By the last corollary, we have $\ulcorner \Phi_\alpha(\delta) \urcorner \in tc$; we can also show that $\mathsf{KF} \vdash \forall \delta [T \ulcorner \Phi_\alpha(\dot{\delta}) \urcorner \leftrightarrow \Phi_\alpha(\delta)]$. Then, by applying Lemma 4.6.4 to the case where $a = \ulcorner \Phi_\alpha \urcorner$, we obtain the next lemma.

Lemma 4.6.7. Let $\alpha < \varepsilon_0$ and $\beta < \varphi_{\varepsilon_0} 0$. Then $\mathsf{KF} \vdash \forall \delta < \bar{\beta} \Phi(\delta)$: i.e.,

$$\mathsf{KF} \vdash \forall \delta < \bar{\beta} \forall C \subset \mathrm{St}_{\mathcal{L}_1^{\mathsf{A}}}\big[\mathsf{ah}_1(\delta, C) \to T\big(h_{\bar{\alpha}}(\dot{\bigwedge} C)\big)\big].$$

Finally, Lemmata 4.6.5 and 4.6.7 yield the above (I). We can similarly obtain (II) as well; in

fact, we need not use the translation like $\mathcal{H}_\alpha$ and can directly show that, for each given $\beta < \Gamma_0$,

$$\mathsf{KF}^* \vdash \forall \delta < \bar{\beta} \forall C \subset \mathrm{St}_{\mathcal{L}_1^{\mathsf{A}}}\big[\mathsf{ah}_1(\delta, C) \to T(\dot{\bigwedge} C)\big];$$

this yields (II).

All the arguments so far applies to the cases of the other iterative compositional truths with no serious modifications. Hence, we finally obtain:

Theorem 4.6.8. $\mathsf{KF}^{(*)}$+Cons $\leq_{\mathcal{I}} \mathsf{KF}^{(*)}$, $\mathsf{FKF}^{(*)}$+Cons $\leq_{\mathcal{I}} \mathsf{FKF}^{(*)}$, and $\mathsf{WKF}^{(*)}$+Cons $\leq_{\mathcal{I}} \mathsf{WKF}^{(*)}$.

4.6.2 Comparison of different self-applicable truths

The main theorem of the present subsection is:

Theorem 4.6.9. $\mathsf{WKF}^{(*)} \leq_{\mathcal{I}} \mathsf{FKF}^{(*)}$ and $\mathsf{FKF}^{(*)} \leq_{\mathcal{I}} \mathsf{KF}^{(*)}$.

This theorem follows from the next two lemmata by similar arguments to the last subsection.

Lemma 4.6.10. Let Γ be a finite set of $\mathcal{L}_1^{\mathsf{A}}$-sentences in which every occurrence of P_1^ξ (for some ξ) is positive. Suppose $\mathsf{AH}_1(\mathcal{C}) \vdash_1^\delta \Gamma, P_1^\eta(t)$, for some ordinal η and a (closed) term t. Then, we have $\mathsf{AH}_n(\mathcal{B}) \vdash_1^\delta \Gamma, P_1^\eta(t)$.

Lemma 4.6.11. Let Γ be a finite set of $\mathcal{L}_1^{\mathsf{A}}$-sentences in which every occurrence of P_1^ξ (for some ξ) is positive. Suppose $\mathsf{AH}_1(\mathcal{B}) \vdash_1^\delta \Gamma, P_1^\eta(t)$, for some ordinal η and a (closed) term t. Then, we have $\mathsf{AH}_n(\mathcal{A}) \vdash_1^\delta \Gamma, P_1^\eta(t)$.

Both are proved by straightforward induction on δ by using the fact that if $\mathsf{AH}_1(\mathcal{C}) \vdash_1^\delta \Delta$ (or $\mathsf{AH}_1^{\mathcal{B}} \vdash_1^\delta$)) for a Δ in which every occurrence of P_1^ξ (for some ξ) is positive, then we can assume that the derivation contains no occurrence of $\neg P_1^{\xi'}$ (for any ξ') and thus no negative fixed-point rule is used therein; therefore, the proofs reduces to showing that, for each Γ satisfying the conditions of

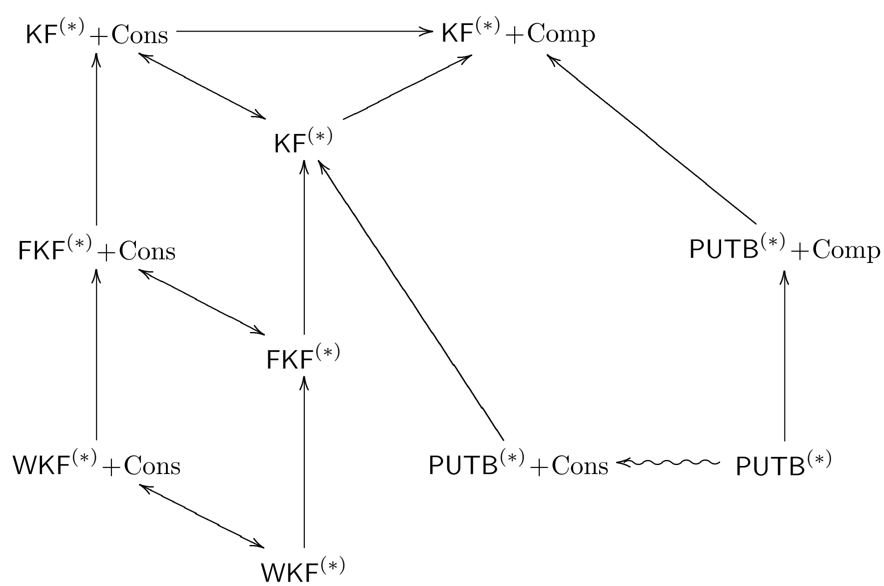
the above lemmata,

$$\mathsf{AH}_1(\mathcal{C}) \overset{\delta}{\underset{1}{\vdash}} \Gamma, \mathcal{C}_1(P_1^\beta, t) \quad \Rightarrow \quad \mathsf{AH}_1(\mathcal{B}) \overset{\delta}{\underset{1}{\vdash}} \Gamma, \mathcal{B}_1(P_1^\beta, t)$$
$$\mathsf{AH}_1(\mathcal{C}) \overset{\delta}{\underset{1}{\vdash}} \Gamma, \mathcal{B}_1(P_1^\beta, t) \quad \Rightarrow \quad \mathsf{AH}_1(\mathcal{B}) \overset{\delta}{\underset{1}{\vdash}} \Gamma, \mathcal{A}_1(P_1^\beta, t);$$

the detail is somehow tedious but straightforward, since, roughly speaking, $\mathcal{C}_1$ implies $\mathcal{B}_1$ and $\mathcal{B}_1$ implies $\mathcal{A}_1$ in PA without using no fixed-point axiom.

4.7 Summary of Chapter 4

The results obtained in §§5–6 is summarized in the next diagram:



This diagram is to be read in the same way as the last diagram except that the curly arrow means that the converse is not yet known to hold.

Unfortunately, as the diagram indicates, two open problems are still left to us: i.e., it isn't yet known whether $\mathsf{PUTB} + \mathrm{Cons} \leq_{\mathcal{I}} \mathsf{PUTB}$ and whether $(\mathsf{PUTB} + \mathrm{Cons})^* \leq_{\mathcal{I}} \mathsf{PUTB}^*$; I conjecture

that both indeed hold but don't know their proofs.

One obvious direction one can take for further studies on inner theory is to investigate the inner theories of other systems of self-applicable truth: e.g., VF or the systems of Friedman and Sheard [23] (their analyses are already given by Cantini [6], Halbach [31] and Leigh and Rathjen [52]). However, it is particularly when the target systems all have the same truth-free consequences that inner theory comparison of systems are of the most significance, since the principal objective of my proposing inner theory comparison consists in the aforementioned motivation of differentiating systems of truth with distinct conceptions by means of a finer-grained criterion than the mere comparison by their truth-free consequences.

Chapter 5

Truth Progressions

5.1 Background

The present chapter studies several systems of the transfinite iteration and autonomous progression of self-applicable truth and determines their proof-theoretic strength.

To my knowledge, Jäger et al. [43] first introduced and studied a transfinitely iterated self-applicable truth. They presented the system of transfinitely iterated Kripke-Feferman truth and gave its proof-theoretic analysis. Then, Strahm [78] determined the proof-theoretic strength of the autonomous progression of Kripke-Feferman truth. the present chapter extends these studies to systems of other kinds of self-applicable truths.

Besides Kripke-Feferman truth, we can consider the autonomous progressions and transfinite iterations of various other conceptions and axiomatizations of self-applicable truth which have already been introduced so far. In the present chapter, we formulate the iterations and autonomous progressions of them and give their proof-theoretic analyses. Specifically, we study those of: Kripke-Feferman truth based on weak Kleene schema axiomatized by WKF; a purely disquotational truth represented by PUTB; Feferman's determinate truth represented by DT; their variants with Fried-

man and Sheard's consistency or completeness axiom; and iterative but non-compositional self-applicable truth of Cantini's system VF. In particular, the system $\mathsf{Aut}(\mathsf{VF})$ of autonomously iterated VF truth has a fairly high proof-theoretic strength compared to the other systems of truth presented so far; indeed, as far as I know, it is the strongest system of truth so far.

The results of the present chapter, particularly on autonomous progressions, bear a close connection to the foundational questions in mathematics. In the rest of the present introductory section, we explain this point.

The notion of autonomous progression of theories was proposed by Kreisel and Feferman. It arose from the following questions: What is implicit in accepting a certain mathematical system S and what ought we to accept, on the same fundamental grounds, once we have made a commitment to S? This, in turn, also tells us what we *can* accept or be justified to accept, once we become justified in accepting a starting system.

One accepts a mathematical system S only when she believes that S possesses a certain favorable property for her interests. For example, if one considers S inconsistent, she doesn't accept S, and, in the usual setting, consistency of S is implicitly accepted in her accepting the system S itself. For another example, one might argue that by accepting S we mean to accept that S is sound, and thus a certain form of reflection principle for S would be implicit in our acceptance of S. Gödel's theorems tell us that neither consistency nor reflection principle of S is derivable within S itself. Hence, according to these views of 'acceptance of system', one implicitly accepts something beyond S in accepting S itself, and, as Feferman wrote, '[Gödel's theorems] point to the possibility of systematically generating larger and larger systems whose acceptability is implicit in acceptance of the starting theory ([17, p.1])'. Autonomous progression is one of the most natural formulations of such a systematic generation for any given view of 'acceptance of system'.

In contrast to this consideration on what we *ought* to accept, the notion of autonomous progression also suggests an answer to the question of what we *can* accept, on the same fundamental

grounds as the starting system S, once we become justified in accepting S. Feferman's approach to predicativism is a typical example of this idea: once a system S is accepted, all the sets definable in S can be justifiably posited and quantification over them are allowed; then we can autonomously iterate this procedure and achieve the system of predicative analysis.

Now, let us focus on one specific but very natural view of 'acceptance of system': namely, the view that by accepting a system S we mean that we accept S as *true*. Then, what does it mean to accept a system as true? We consider the following interpretation: once we accept a system S_0, we implicitly accept it as true; we add a new predicate T to S_0 for expressing its truth (a new predicate is needed due to Tarski's undefinability theorem); since we have accepted S_0 as true, the axioms expressing 'T represents the truth predicate for the language of S_0' and 'S_0 is true (in terms of the truth predicate T)' ought to and can be accepted. In this way, we obtain and accept a new system S_1. Then, we can iterate this procedure of generating (and accepting) $\mathsf{S}_0, \mathsf{S}_1, \ldots, \mathsf{S}_n, \ldots$ even for transfinitely many steps (by taking the union at a limit level) along autonomously obtained well-orderings.

This progression depends on how truth is axiomatized. It is a highly profound and difficult question: By what axioms and rules the mathematical truth is to be characterized? The best known proposal is the Tarskian conception of truth, and the ordinal of its autonomous progression is known to be the Feferman-Schütte ordinal Γ_0; see [17] or [24]. However, the Tarskian truth is typed and has been criticized by philosophers and logicians as not capturing the essential self-applicable character of the notion of truth, and various approaches have been attempted for axiomatizing self-applicable truth. It is of course not easy at all to pick out the 'genuine' axiomatization of self-applicable truth. the present chapter will examine the autonomous progressions (and iterations) of some possible options. We expect that our results, together with the cited works of Jäger et al. and Strahm, shed more light on the study of self-applicable truth.

5.2 Notation for Transfinitely Iterated Truths

For the subsequent arguments, we need slightly different notations from what has been used so far. First of all, for technical simplicity, I will exclude '$\to$' from the primitive logical connective and assume that they are only '$\neg$' and '$\wedge$' throughout the present chapter.

Let $\mathcal{L}_{\mathsf{PA}}$ denote the language of first order arithmetic with symbols for all primitive recursive (p.r.) functions; we assume that Peano Arithmetic PA contains all the axioms defining them. Since truth systems lack a second-order parameter, we add a new unary predicate U, which acts like a free set variable, so as to deal with autonomy rules and proof-theoretic ordinals. Our base language $\mathcal{L}_0$ is $\mathcal{L}_{\mathsf{PA}} \cup \{U\}$. We use '$\equiv$' for (meta-) identity between syntactical objects such as formulae (instead of '$=$').

Given any language $\mathcal{L}'$ ($\supset \mathcal{L}_{\mathsf{PA}}$), *full-induction* for $\mathcal{L}'$ is the schema:

$$\phi(0) \wedge \forall x\big(\phi(x) \to \phi(x+1)\big) \to \forall x \phi(x), \text{ for each } \mathcal{L}'\text{-formula } \phi.$$

Our base system is Peano Arithmetic over $\mathcal{L}_0$ and we denote it by $\mathsf{PA}(U)$: i.e., PA plus full-induction for $\mathcal{L}_0$.[1] Given a formula $\phi(x)$ of any language $\mathcal{L}'$ and a p.r. ordering $\prec$, *transfinite induction* along $\prec$ with respect to $\phi(x)$, written $\mathrm{TI}(\prec;\phi)$, is defined by

$$\forall z[(\forall y \prec z)\phi(y) \to \phi(z)] \to \forall x \in \mathrm{fd}(\prec)\phi(x),$$

where $\mathrm{fd}(\prec)$ denotes the field of the relation $\prec$. Then, we define the *proof-theoretic ordinal* $|\mathsf{S}|$ of

[1] We adopt PA in our framework as the base system. One may choose another setting. For example, Cantini [7] and Kahle [46] rather take applicative theories with natural numbers as their base systems; a certain transfinite iteration of self-applicable truth in such a framework is studied in [46].

a system S over any language including $\mathcal{L}_0$ as

$$\sup\{otyp(\prec) \mid \prec \text{ is a p.r. ordering and } \mathsf{S} \vdash \mathrm{TI}(\prec; U)\},$$

where $otyp(\prec)$ denotes the order-type of $\prec$. For example, $|\mathsf{PA}(U)| = \varepsilon_0$. Given any language $\mathcal{L}'$, $\mathrm{TI}_{\mathcal{L}'}(\prec)$ denotes the schema $\mathrm{TI}_{\mathcal{L}'}(\prec; \phi)$ for all $\phi \in \mathcal{L}'$. In all the systems we consider in this chapter, $\mathrm{TI}(\prec; U)$ entails $\mathrm{TI}_{\mathcal{L}_0}(\prec)$.

Given a p.r. ordering $\prec$, $\mathcal{L}^{\prec}$ denotes the language $\mathcal{L}_0 \cup \{T^{\prec}\}$, where $T^{\prec}$ is a new *binary* predicate symbol. We write $T^{\prec}_a(b)$ for $T^{\prec}(b, a)$. For $a \in \mathrm{fd}(\prec)$, $T^{\prec}_a$ is meant to express the a-times iterated truth (or 'true at the a-th level') along $\prec$. Given $a \in \mathrm{fd}(\prec)$, $\mathcal{L}^{\prec}_a$ denotes the sublanguage of $\mathcal{L}^{\prec}$ obtained by restricting atoms of the form $T^{\prec}_s t$ to closed terms s with $s^{\mathbb{N}} \preceq a^{\mathbb{N}}$ (t can be an arbitrary term). The language $\mathcal{L}$ for systems of autonomous progression is defined as $\bigcup\{\mathcal{L}^{\prec} \mid \prec \text{ is a p.r. ordering}\}$.

In order to formulate axiomatic systems of truth, we presuppose a standard Gödelization of $\mathcal{L}$. Each syntactical expression e from the vocabulary of $\mathcal{L}$ is assigned a fixed Gödel number denoted by $\#e$. Let $\ulcorner e \urcorner$ denote the numeral whose value is $\#e$.

First, we need the representations of some basic syntactical relations. $\mathrm{Tm}(x)$, $\mathrm{CT}(x)$ and $\mathrm{Var}(x)$ are representations of the sets of all terms, closed terms and variables respectively. $\mathrm{AtFml}_0(x)$ ($\mathrm{AtSt}_0(x)$), $\mathrm{Fml}_0(x)$ ($\mathrm{St}_0(x)$) and $\mathrm{For}_0(x)$ respectively represent the sets of all atomic formulae (atomic sentences), formulae (sentences) and formulae with at most one free variable of $\mathcal{L}_0$; for each p.r. ordering $\prec$, $\mathrm{AtFml}^{\prec}$ ($\mathrm{AtSt}^{\prec}$), $\mathrm{Fml}^{\prec}$ ($\mathrm{St}^{\prec}$) and $\mathrm{For}^{\prec}$ represent those of $\mathcal{L}^{\prec}$; given $a \in \mathrm{fd}(\prec)$, $\mathrm{AtFml}^{\prec}_a$ ($\mathrm{AtSt}^{\prec}_a$), $\mathrm{Fml}^{\prec}_a$ ($\mathrm{St}^{\prec}_a$) and $\mathrm{For}^{\prec}_a$ are, respectively, the representations of those of $\mathcal{L}^{\prec}_a$.

Next, we introduce the representations of the following syntactical operations: $\mathrm{nm}(x)$ ('the x-th numeral'); $\mathrm{val}(x)$ ('the value of the closed term x') $\dot{R}\vec{x}$ ('the atomic formula obtained by applying terms represented by $\vec{x}$ to R') for each predicate R; $\dot{\neg} x$ ('the negation of the formula x');

$x \dot{\wedge} y$ ('the conjunction of the formulae x and y'); $\dot{\vee}$ for disjunction defined analogously; $\dot{\forall} z.x$ ('the universal quantification of the formula x w.r.t. the variable z'); $\dot{\exists}$ for existential quantification defined analogously. We will usually write $\dot{x}$ and x° respectively for $\mathrm{nm}(x)$ and $\mathrm{val}(x)$ for saving space. For a code a of a term or formula and $c_1, \ldots, c_n \in \mathrm{Tm}$, $\mathrm{sb}(a; c_1, \ldots, c_n)$ denotes (the code of) the term or formula obtained from a by substituting the term c_i for the i-th free variable (in some fixed enumeration of variables) in a for each $i \leq n$; e.g., $\mathrm{sb}(\ulcorner \phi(x_0, x_1) \urcorner; \ulcorner s \urcorner, \ulcorner t \urcorner) = \ulcorner \phi(s,t) \urcorner$ for terms s and t. Then, we set $a(\vec{x}) := \mathrm{sb}(a; \vec{\dot{x}})$, and, given a term or formula e, we write $\ulcorner e(\vec{x}) \urcorner$ for $\ulcorner e \urcorner(\vec{x})$ $(= \mathrm{sb}(\ulcorner e \urcorner; \vec{\dot{x}}))$; we also write $\ulcorner e(\vec{a}) \urcorner$ for $\mathrm{sb}(\ulcorner e(\vec{v}) \urcorner; \vec{a})$.[2] When we need not explicitly specify a variable, we often omit it; e.g., $\ulcorner T_a \dot{b}(\cdot) \urcorner$ means $\ulcorner T_a \dot{b}(w) \urcorner$ for a fresh variable w. We define the *falsity predicate* $F_a^\prec x$ (at the a-th level) as $T_a^\prec(\dot{\neg} x)$; then $\dot{F}$ has its obvious meaning.

All the above syntactical relations and operations are primitive recursive except for $\mathrm{val}(x)$, $\mathrm{AtFml}_a^\prec$, $\mathrm{AtSt}_a^\prec$, $\mathrm{Fml}_a^\prec$, $\mathrm{St}_a^\prec$ and $\mathrm{For}_a^\prec$. Hence, we assume that they are all represented by their corresponding function symbols in $\mathcal{L}_0$, since $\mathcal{L}_0$ possesses symbols for all p.r. functions. The value function $\mathrm{val}(x)$ for $\mathcal{L}_0$-terms is recursive and definable in PA but not p.r., since we could otherwise construct a p.r. evaluation function of all the n-ary p.r. functions (for each n) by means of $\mathrm{val}(x)$; $\mathrm{AtFml}_a^\prec$, $\mathrm{AtSt}_a^\prec$, $\mathrm{Fml}_a^\prec$, $\mathrm{St}_a^\prec$ and $\mathrm{For}_a^\prec$ are also recursive and PA-definable but not p.r., since they need $\mathrm{val}(x)$ in their standard definitions. Hence, $\mathcal{L}_0$ does not contain primitive symbols for these six notions. Let $\mathsf{PA}^+(U)$ be the definitional expansion of $\mathsf{PA}(U)$ augmented by the symbols and defining axioms for them and let $\mathcal{L}_0^+$ be its language. We can either work within this definitional expansion or simply regard them as abbreviations of suitable $\mathcal{L}_0$-formulae representing them.

Remark 8. There are some subtle issues to be noted concerning coding particularly when we work within definitional expansions. First, our coding was fixed before taking definitional expansions. Hence, definitionally introduced symbols are not assigned their own codings. Of course, we can

[2] To avoid possible confusion, we notice that $\mathrm{sb}(a; \vec{c})$ is denoted by $a(\vec{c})$ in [43] and $a(\vec{c})$ in our notation corresponds to their $a(\vec{\dot{c}})$, but $\ulcorner \phi(\vec{x}) \urcorner$ has the same meaning for a formula ϕ.

redefine coding so as to accommodate the new symbols, but then $\mathrm{val}(x)$ is no longer the value function in this new coding schema. Thus, even when we work within $\mathcal{L}_0^+$ or any further definitional expansion $\mathcal{L}'$, the codes of $\mathcal{L}'$-expressions should be regarded as the codes of corresponding $\mathcal{L}_0$-expressions with the defining formulae of the new symbols. Second, truth systems sometimes behave in unexpected ways on definitionally introduced functions. For example, let f be a definable function and $\phi(x,y)$ be its defining formula. Consider a sentence $T_s^{\prec}\ulcorner T_s^{\prec} f(\dot{x})\urcorner$ for a closed term s with $s^{\mathbb{N}} \in \mathrm{fd}(\prec)$. Then, one might expect that it is equivalent to $T_s\big(f(x)\big)$ due to self-applicability of truth. However, $T_s^{\prec}\ulcorner T_s^{\prec} f(\dot{x})\urcorner$ indeed means, say, $T_s^{\prec}\ulcorner \forall y[\phi(\dot{x},y) \to T_s^{\prec}(y)\urcorner$. Then, whether they are equivalent depends on the inner logic (or the compositional axioms) of the system in question. As a matter of fact, this equivalence fails in $\mathsf{WKF}_{\prec}$ (+Cons); nonetheless, as we shall see in §6, this is no obstacle to well-ordering proofs and has no essential effect on the proof-theoretic strength. In contrast, they are equivalent and definable functions can generally be treated as if they were primitive symbols within all the other systems in the present chapter; this fact is implicitly made use of particularly in Lemmata 5.4.2, 5.5.2 and 5.7.4.

5.3 Kripke-Feferman Truth

In the present section, we introduce the transfinite iteration and autonomous progression of Kripke-Feferman truth and review the results of Jäger et al. [43] and Strahm [78] on them.

Definition 5.3.1. Given a p.r. ordering $\prec$, the system $\mathsf{KF}_{\prec}$ over $\mathcal{L}^{\prec}$ consists of $\mathsf{PA}(U)$ plus full-induction for $\mathcal{L}^{\prec}$, $\mathrm{TI}_{\mathcal{L}^{\prec}}(\prec)$ and the following axioms:

K1$_{\prec}$ $\forall a \in \mathrm{fd}(\prec)\forall \vec{x} \in \mathrm{CT}\big[\big(T_a^{\prec}(\dot{R}\vec{x}) \leftrightarrow R\vec{x}^{\circ}\big) \wedge \big(F_a^{\prec}(\dot{R}\vec{x}) \leftrightarrow \neg R\vec{x}^{\circ}\big]$,

for each predicate symbol R of $\mathcal{L}_0$

K2$_{\prec}$ $\forall a \in \mathrm{fd}(\prec)\forall x, b \in \mathrm{CT}$

$\big[b^{\circ} \prec a \to \big(T_a^{\prec}(\dot{T}_b^{\prec} x) \leftrightarrow T_{b^{\circ}}^{\prec}(x^{\circ})\big) \wedge \big(F_a^{\prec}(\dot{T}_b^{\prec} x) \leftrightarrow \neg T_{b^{\circ}}^{\prec}(x^{\circ})\big)\big]$

$\mathbf{K3}_\prec$ $\forall x, a \in \mathrm{CT}\big[a^\circ \in \mathrm{fd}(\prec) \to \big(T^\prec_{a^\circ}(\dot{T}^\prec_a x) \leftrightarrow T^\prec_{a^\circ} x^\circ\big) \wedge \big(F^\prec_{a^\circ}(\dot{T}^\prec_a x) \leftrightarrow F^\prec_{a^\circ} x^\circ\big)\big]$

$\mathbf{K4}_\prec$ $\forall a \in \mathrm{fd}(\prec) \forall x \in \mathrm{St}^\prec_a \big[T^\prec_a(\dot{\neg}\dot{\neg} x) \leftrightarrow T^\prec_a x\big]$

$\mathbf{K5}_\prec$ $\forall a \in \mathrm{fd}(\prec) \forall x, y \in \mathrm{St}^\prec_a \big[\big(T^\prec_a(x \dot{\wedge} y) \leftrightarrow T^\prec_a x \wedge T^\prec_a y\big) \wedge \big(F^\prec_a(x \dot{\wedge} y) \leftrightarrow F^\prec_a x \vee F^\prec_a y\big)\big]$

$\mathbf{K6}_\prec$ $\forall a \in \mathrm{fd}(\prec) \forall z, x$

$\big[\dot{\forall} z.x \in \mathrm{St}^\prec_a \to \big(T^\prec_a(\dot{\forall} z.x) \leftrightarrow \forall y T^\prec_a x(y)\big) \wedge \big(F^\prec_a(\dot{\forall} z.x) \leftrightarrow \exists y F^\prec_a x(y)\big)\big]$

The system $\mathsf{Aut}(\mathsf{KF})$ over $\mathcal{L}$ consists of $\mathsf{PA}(U)$, full-induction for $\mathcal{L}$, and the following two rules: for each p.r. linear ordering (provably in, say, $\mathsf{PA}(U)$),

$$\frac{\mathrm{TI}(\prec; U)}{\mathbf{K1}_\prec \wedge \ldots \wedge \mathbf{K6}_\prec} \quad \text{and} \quad \frac{\mathrm{TI}(\prec; U)}{\mathrm{TI}(\prec; \psi)} \ \text{for each } \mathcal{L}\text{-formula } \psi.$$

The right-hand rule corresponds to the Bar rule, BR. Indeed, $\mathsf{KF}_\prec$ is inconsistent for some orderings $\prec$, but this autonomy rule blocks inconsistency by restricting $\prec$ to well-orderings.

Ordinals up to the Feferman-Schütte ordinal Γ_0 are not sufficient for ordinal analyses of the systems we will consider. Recall that Γ_0 is the least ordinal (> 0) closed under the binary Veblen function φ. We introduce the *ternary Veblen function* φ by a simple generalization of the binary one: $\varphi 0 \alpha \beta = \varphi_\alpha \beta$; $\varphi \alpha 0 \gamma$ for $\alpha > 0$ is the γ-th strongly critical ordinal with respect to $\lambda \xi, \eta.\ \varphi \alpha' \xi \eta$ for $\alpha' < \alpha$; $\varphi \alpha \beta \gamma$ for $\alpha, \beta > 0$ is the γ-th common fixed-point of $\lambda \xi. \varphi \alpha \beta' \xi$ for $\beta' < \beta$. Let Λ_3 be the least ordinal > 0 closed under the ternary Veblen function. Throughout §§3-7, we fix a natural ordinal notation system up to a sufficiently large ordinal, say, Λ_3, and let $\lhd$ denote the corresponding well-ordering. Given a closed term $t \in \mathrm{fd}(\lhd)$, $|t|$ denotes the unique ordinal associated to t by this notation system. For $\alpha \in \mathrm{fd}(\lhd)$, let $\lhd \restriction_\alpha$ be $\{\langle \gamma, \beta \rangle \mid \gamma \lhd \beta \lhd \alpha\}$. We write KF_α for $\mathsf{KF}_{\lhd \restriction_\alpha}$ for $\alpha \in \mathrm{fd}(\lhd)$ (and similar for other systems). For a language $\mathcal{L}'$, we write $\mathrm{TI}_{\mathcal{L}'}(\alpha)$ for $\mathrm{TI}_{\mathcal{L}'}(\lhd \restriction_\alpha)$.

Jägar et al. [43] determined the proof-theoretic ordinal of the system $\widehat{\mathsf{ID}}_\alpha$ of α-times iterated fixed-points. Let $\mathcal{L}_0^{(+)}(R, Q)$ be $\mathcal{L}_0^{(+)} \cup \{R, Q\}$ where R and Q are fresh unary predicate symbols.

An $\mathcal{L}_0^{(+)}(R,Q)$-formula $\mathcal{A}(R,Q,x,y)$ with at most x and y free is called an *inductive operator form*, when R occurs only positively in $\mathcal{A}$. The language $\mathcal{L}_{\rm fix}$ of $\widehat{\mathsf{ID}}_\prec$ for a given $\prec$ is defined as $\mathcal{L}_0$ plus *unary* predicates $P^{\mathcal{A}}$ associated to each inductive operator form $\mathcal{A}(R,Q,x,y) \in \mathcal{L}_0(R,Q)$. We write $P^{\mathcal{A}}_x(y)$ for $P^{\mathcal{A}}(\langle y,x\rangle)$ and $P^{\mathcal{A}}_{\prec x}(z)$ for $(z)_1 \prec x \wedge P^{\mathcal{A}}(z)$, where $\langle\cdot,\cdot\rangle$ is a p.r. paring function. We assume that the paring function is bijective from $\mathbb{N}\times\mathbb{N}$ to $\mathbb{N}$ and we let $(\cdot)_0$ and $(\cdot)_1$ be the corresponding projections; thus, in contrast to [43], we do not need the clause $z \in \mathrm{Pair}$ ('z is a code of a pair') in the definition of $P^{\mathcal{A}}_{\prec x}(z)$. Then, the system $\widehat{\mathsf{ID}}_\prec$ comprises $\mathsf{PA}(U)$, full-induction for $\mathcal{L}_{\rm fix}$, $\mathrm{TI}_{\mathcal{L}_{\rm fix}}(\prec)$, and the fixed-point axiom schema: for each inductive operator form $\mathcal{A}$,

$$\forall a \in \mathrm{fd}(\prec)\forall x[P^{\mathcal{A}}_a(x) \leftrightarrow \mathcal{A}(P^{\mathcal{A}}_a, P^{\mathcal{A}}_{\prec a}, x, a)].$$

In giving the lower bound of $|\widehat{\mathsf{ID}}_\alpha (= \widehat{\mathsf{ID}}_{\lhd\uparrow\alpha})|$, Jäger et al. used an intermediate system SRT_α, which is embeddable in ID_α. This SRT_α is essentially the same system as KF_α.[3] It is observed from their proof that KF_α and $\widehat{\mathsf{ID}}_\alpha$ have the same proof-theoretic ordinal:

Theorem 5.3.2 (Jäger et al. [43])**.** Given an ordinal δ, $\varepsilon(\delta)$ denotes the least epsilon number greater than δ, and we define $(\delta|m)$ inductively by $(\delta|0) = \varepsilon(\delta)$ and $(\delta|k+1) = \varphi_{(\delta|k)}0$. Let an ordinal α be such that

$$\alpha = \omega^{1+\alpha_n} + \omega^{1+\alpha_{n-1}} + \cdots + \omega^{1+\alpha_1} + m \quad (\alpha_n \geq \cdots \geq \alpha_1). \tag{5.1}$$

Then, $|\widehat{\mathsf{ID}}_\alpha| = |\mathsf{KF}_\alpha| = \varphi 1\alpha_n(\varphi 1\alpha_{n-1}(\cdots\varphi 1\alpha_1(\alpha|m))\cdots)$.

Strahm [78] introduced the system $\mathsf{Aut}(\widehat{\mathsf{ID}})$ for autonomous fixed-point progression, and showed that its proof-theoretic ordinal is $\varphi 200$. His proof can be directly applied to $\mathsf{Aut}(\mathsf{KF})$ and thus yields the following theorem:

[3] The acronym SRT stands for 'self-reflective truth'. Because all the truths we consider in the present chapter are self-reflective, we rename it in order to contrast it against other conceptions of truth.

Theorem 5.3.3 (Strahm [78]). $|\mathsf{Aut}(\mathsf{KF})| = |\mathsf{Aut}(\widehat{\mathsf{ID}})| = \varphi 200$.

Remark 9. Feferman's original motivation for developing the notion of *reflective closure*, which is what truth theorists now call KF, might be incompatible with its (autonomous) iteration. Feferman's principal motivation behind it was to suggest a 'more realistic and perspicuous' (finite) alternative to autonomous progression for approximating or finding out what is implicit in accepting a mathematical system. In the present chapter, we regard KF not as means of reflective closure but as one axiomatization of (self-applicable) truth.

5.4 Consistency and Completeness Axioms

Friedman and Sheard [23] introduced two special axioms for self-applicable truth: the consistency axiom Cons (expressing 'no sentence is both true and false') and completeness axiom Comp ('every sentence is either true or false'). Cantini [5] then considered adding them to KF and studied the resulting systems.[4] Although neither of them is derivable in KF and adding both to KF leads to a contradiction, either (but not both) of them can be consistently and even conservatively (for $\mathcal{L}_0$) added to KF.

In this section, we study consistency and completeness axioms for iterated truth: for a given p.r. ordering $\prec$,

$$\begin{aligned} \mathrm{Cons}_{\prec} &:\equiv \forall a \in \mathrm{fd}(\prec) \forall x \in \mathrm{St}_a^{\prec}\ \neg(T_a^{\prec} x \wedge F_a^{\prec} x) \\ \mathrm{Comp}_{\prec} &:\equiv \forall a \in \mathrm{fd}(\prec) \forall x \in \mathrm{St}_a^{\prec}\ (T_a^{\prec} x \vee F_a^{\prec} x). \end{aligned}$$

As in the last section, for $\alpha \in \mathrm{fd}(\lhd)$, we write Cons_α and Comp_α for $\mathrm{Cons}_{\lhd\restriction_\alpha}$ and $\mathrm{Comp}_{\lhd\restriction_\alpha}$ respectively. Throughout this section, we write $\gamma < \beta$ for $\gamma \lhd \beta$ for readability and drop the

[4] Feferman [17] also considered the consistency axiom which is denoted by *Disj* there.

superscript $\triangleleft\restriction_\alpha$ whenever there is no worry of confusion. We shall see that Cons_α or Comp_α can also be conservatively added to KF_α.

Definition 5.4.1. $\mathsf{Aut}(\mathsf{KF}+\mathrm{Cons})$ and $\mathsf{Aut}(\mathsf{KF}+\mathrm{Comp})$ over $\mathcal{L}$ are defined as $\mathsf{Aut}(\mathsf{KF})$ plus

$$\frac{\mathrm{TI}(\prec;U)}{\mathrm{Cons}_\prec} \quad \text{or} \quad \frac{\mathrm{TI}(\prec;U)}{\mathrm{Comp}_\prec},$$

for each p.r. linear ordering $\prec$, respectively.

The goal of the present section is to determine the proof-theoretic ordinals of $\mathsf{KF}_\alpha+\mathrm{Cons}_\alpha$, $\mathsf{KF}_\alpha+\mathrm{Comp}_\alpha$, $\mathsf{Aut}(\mathsf{KF}+\mathrm{Cons})$, and $\mathsf{Aut}(\mathsf{KF}+\mathrm{Comp})$. Cantini [5] showed that $\mathsf{KF}+\mathrm{Cons}$ and $\mathsf{KF}+\mathrm{Comp}$ are syntactically embeddable in each other; by *syntactical embedding*, we mean relative interpretation which keeps the vocabulary of the base language unchanged. As the next lemma shows, the same applies to $\mathsf{KF}_\prec$ and thus it suffices for our goal to show $|\mathsf{KF}_\alpha+\mathrm{Cons}_\alpha|\leq|\mathsf{KF}_\alpha|$ via a suitable cut-elimination procedure.

Lemma 5.4.2. $\mathsf{KF}_\prec+\mathrm{Cons}_\prec$ and $\mathsf{KF}_\prec+\mathrm{Comp}_\prec$ are mutually syntactically embeddable.

Proof. We define in $\mathsf{PA}(U)$ (by the recursion theorem or diagonalization) a function f by:

$$f(x):=\begin{cases} x & \text{if } x\in\mathrm{AtFml}_0\\ \ulcorner\neg F^\prec_z(fy)\urcorner\ \big(\text{i.e., } \mathrm{sb}(\ulcorner\neg F(f(\cdot),\cdot)\urcorner;y,z)\big) & \text{if } x=T^\prec_z y \text{ for } y,z\in\mathrm{Tm}\\ \dot\neg f(y)\ \ (f(y)\dot\wedge f(z), \text{ resp.}) & \text{if } x=\dot\neg y \text{ (or } x=y\dot\wedge z)\\ \dot\forall z.f(y) & \text{if } x=\dot\forall z.y; \end{cases}$$

otherwise, $f(x)=0$. The embedding $\mathcal{F}$ for both directions is given by $T^\prec_z x\mapsto\neg F^\prec_z f(x)$. We only demonstrate that $\mathbf{K2}_\prec$ is preserved by $\mathcal{F}$: for $b,x\in\mathrm{CT}$ with $b^\circ\prec a\in\mathrm{fd}(\prec)$:

$$\mathcal{F}(T^\prec_a T^\prec_{\dot b}x)\equiv\neg F^\prec_a\ulcorner\neg F^\prec_b(fx)\urcorner\leftrightarrow\neg T^\prec_a\ulcorner F^\prec_b(fx)\urcorner\leftrightarrow\neg F^\prec_b f(x^\circ)\equiv\mathcal{F}(T^\prec_b x^\circ)$$

$$\mathcal{F}(F^\prec_a T^\prec_{\dot b}x)\equiv\neg F^\prec_a\dot\neg\ulcorner\neg F^\prec_b(fx)\urcorner\leftrightarrow\neg F^\prec_a\ulcorner F^\prec_b(fx)\urcorner\leftrightarrow\neg\neg F^\prec_b f(x^\circ)\equiv\mathcal{F}(\neg T^\prec_b x^\circ).$$

The other cases are straightforward. □

We assume familiarity with the cut-elimination procedure of $\widehat{\mathsf{ID}}_\alpha$ in [43] and use the same notations as there. For technical reasons, we explicitly work with the definitional expansion $\mathsf{PA}^+(U)$; thus, in particular, $\widehat{\mathsf{ID}}_\alpha$ (over $\mathsf{PA}^+(U)$) possesses $\mathrm{val}(x)$ and $\mathrm{St}_z^\prec$ as primitive symbols. In general, $\mathsf{KF}_\prec$ can be embedded in $\widehat{\mathsf{ID}}_\prec$ for any $\prec$; there exists an inductive operator form $\mathcal{A}$ such that if KF_α $(= \mathsf{SRT}_\alpha) \vdash \phi$ then $\widehat{\mathsf{ID}}_\alpha \vdash \phi[P^{\mathcal{A}}/T]$, where $\phi[P^{\mathcal{A}}/T]$ is the result of substituting $P^{\mathcal{A}}_s t$ for $T_s t$ in ϕ for each terms s and t. We can assume that $\widehat{\mathsf{ID}}_\alpha \vdash \phi[P^{\mathcal{A}}/T]$ above is obtained by a derivation in which only the vocabulary from $\mathcal{L}_0^+ \cup \{P^{\mathcal{A}}\}$ is used and no other atomic of the form $P^{\mathcal{B}}$ $(\mathcal{B} \not\equiv \mathcal{A})$ occurs. For the subsequent arguments, we precisely describe one such $\mathcal{A}$ and fix it. We take the following $\mathcal{A}(R, Q, x, a)$:

$$
\begin{aligned}
& x \in \mathrm{St}_a \wedge \forall y, z \in \mathrm{CT}[x = y \dot{=} z \to y^\circ = z^\circ] \wedge \forall y, z \in \mathrm{CT}[x = y \dot{\neq} z \to y^\circ \neq z^\circ] \\
& \wedge \forall y \in \mathrm{CT}[(x = \dot{U} y) \to U(y^\circ)] \wedge \forall y \in \mathrm{CT}[(x = \dot{\neg} \dot{U} y) \to \neg U(y^\circ)] \\
& \wedge \forall b, y \in \mathrm{CT}[(b^\circ < a \wedge x = \dot{T}_b y) \to \langle y^\circ, b^\circ \rangle \in Q] \\
& \wedge \forall b, y \in \mathrm{CT}[(b^\circ < a \wedge x = \dot{\neg} \dot{T}_b y) \to \langle y^\circ, b^\circ \rangle \notin Q] \\
& \wedge \forall b, y \in \mathrm{CT}[(b^\circ = a \wedge x = \dot{T}_b y) \to y^\circ \in R] \\
& \wedge \forall b, y \in \mathrm{CT}[(b^\circ = a \wedge x = \dot{\neg} \dot{T}_b y) \to \dot{\neg} y^\circ \in R] \\
& \wedge \forall y \in \mathrm{St}_a[(x = \dot{\neg}\dot{\neg} y) \to y \in R] \wedge \forall y, z \in \mathrm{St}_a[(x = y \dot{\wedge} z) \to Ry \wedge Rz] \\
& \wedge \forall y, z \in \mathrm{St}_a\big[(x = \dot{\neg}(y \dot{\wedge} z)) \to \big(\dot{\neg} y \in R \vee \dot{\neg} z \in R\big)\big] \\
& \wedge \forall y, z[\mathrm{St}_a(\dot{\forall} y.z) \wedge x = \dot{\forall} y.z \to \forall w(z(w) \in R)] \\
& \wedge \forall y, z[\mathrm{St}_a(\dot{\forall} y.z) \wedge x = \dot{\neg}\dot{\forall} y.z \to \exists w(\dot{\neg} z(w) \in R)].
\end{aligned}
$$

We use the same semi-formal system H_α for $\widehat{\mathsf{ID}}_\alpha$ as in [43, p.62-3] and follow the same notations; but we need to make slight modifications to accommodate the new symbols in $\mathcal{L}_0^+$ and we postulate

all the true $\mathcal{L}_0^+$-literals (e.g., $\mathrm{val}(\ulcorner 1+1 \urcorner) = 2$) as axioms; cf. the definition of $\mathsf{AH}_{\alpha'}$ below.[5] Since we focus only on the above specific $\mathcal{A}$, we can assume that the language $\mathcal{L}_{\mathrm{fix}}^{\alpha}$ (written as $\mathcal{L}_\alpha$ in [43]) of H_α extends $\mathcal{L}_0^+$ only by unary symbols $P_\beta^{\mathcal{A}}$ ($\beta < \alpha$) and $P_{<\gamma}^{\mathcal{A}}$ ($\gamma \leq \alpha$). Recall that $\mathsf{H}_\alpha \vdash_\star^\delta \Gamma$ (or $\mathsf{H}_\alpha \vdash_\star^{<\delta} \Gamma$) means that there is a derivation of Γ in H_α with the length δ ($< \delta$, resp.) in which all its cut formulae are of the forms $P_\beta^{\mathcal{A}} t$ and $\neg P_\beta^{\mathcal{A}}$ ($\beta < \alpha$). As was shown in [43] (Proposition 15), if $\widehat{\mathsf{ID}}_\alpha \vdash A$ for an $\mathcal{L}_{\mathrm{fix}}$-sentence A then $\mathsf{H}_\alpha \vdash_\star^{<\varepsilon(\alpha)} A^\alpha$, where the translation A^α is obtained from A by replacing $P^{\mathcal{A}}$ by $P_{<\alpha}^{\mathcal{A}}$. In particular, if $\mathsf{KF}_\alpha + \mathrm{Cons}_\alpha \vdash \sigma$ for some $\mathcal{L}_0^+$-sentence σ, then we have $\widehat{\mathsf{ID}}_\alpha \vdash \mathrm{Cons}_\alpha[P^{\mathcal{A}}/T] \to \sigma$ and thus $\mathsf{H}_\alpha \vdash_\star^{<\varepsilon(\alpha)} (\neg \mathrm{Cons}_\alpha[P^{\mathcal{A}}/T])^\alpha, \sigma$. In what follows, we just identify Cons_α with $(\mathrm{Cons}_\alpha[P^{\mathcal{A}}/T])^\alpha$ and omit the superscript $\mathcal{A}$ for simplicity.

Lemma 5.4.3. Let $\gamma < \beta \leq \alpha$ and s and t be closed terms with $(t)_0^{\mathbb{N}} = s^{\mathbb{N}}$ and $|(t)_1| = \gamma$. If $\mathsf{H}_\alpha \vdash^\delta \phi, \Gamma$ with arbitrary cut-rank, then $\mathsf{H}_\alpha \vdash^{1+\delta} \phi[P_\gamma(s)/P_{<\beta}(t)], \Gamma$ with the same cut-rank.

Proof. Induction on δ. There are two essential cases. One is the base step where ϕ is obtained by the Axiom I ([43, p.62]) and $\phi \equiv P_{<\beta}(t)$ or $\phi \equiv \neg P_{<\beta}(t)$. We illustrate the former case; then, $\Gamma = \Gamma' \cup \{\neg P_{<\beta}(r)\}$ for some Γ' and $r^{\mathbb{N}} = t^{\mathbb{N}}$. The same axiom yields $\mathsf{H}_\alpha \vdash^0 \Gamma', P_\gamma(s), \neg P_\gamma((r)_0)$. Then we obtain $\mathsf{H}_\alpha \vdash^1 \Gamma', P_\gamma(s), \neg P_{<\beta}(r)$ by the Fixed-point rule III ([43, p.63]). The other essential case is where ϕ is obtained by the Fixed-point rule III. In this case, the premise is numerically equivalent to the claimed sequent itself. □

Lemma 5.4.4. Let $\gamma \leq \beta \leq \alpha$ and t be a closed term with $|(t)_1| < \gamma$. Suppose $\mathsf{H}_\alpha \vdash^\delta \Gamma, \phi$. Then, $\mathsf{H}_\alpha \vdash^{2+\delta} \Gamma, \phi[P_{<\gamma}(t)/P_{<\beta}(t)]$ (with the same cut-rank).

Proof. Induction on δ. The essential cases are the same. For the first case, we apply the Fixed-point rule III twice to a suitable instance of the Axiom I. For the second, we obtain $(\neg)P_{<\gamma}(t)$ instead of $(\neg)P_{<\beta}(t)$ from the premise by the same rule. □

[5] This change results in a slight strengthening of the ordinary ω-arithmetic, but, even with this change, all the arguments in [43, §6] can be carried out in the same way and cut-elimination of H_α properly gives the upper bound of $\widehat{\mathsf{ID}}_\alpha$; cf. the proof of Theorem 1.3.6 of [64].

We make use of the technique of asymmetric interpretation of H_α. Given a successor $\alpha' = \alpha+1$, we introduce a Tait-style semi-formal system $\mathsf{AH}_{\alpha'}$. The language $\mathcal{L}_{\mathsf{A}}^{\alpha'}$ ($\supset \mathcal{L}_{\mathrm{fix}}^{\alpha}$) of $\mathsf{AH}_{\alpha'}$ consists of the sentences (in Tait-style) of $\mathcal{L}_0^+ \cup \{P_\xi \mid \xi < \alpha\} \cup \{P_{<\xi} \mid \xi \leq \alpha\} \cup \{P_\alpha^\eta \mid \eta\}$, where η ranges over ordinals up to a sufficiently large ordinal, e.g., Λ_3; namely, the predicate P_α of the highest level is further typed by ordinals. Note that, in contrast to $\mathcal{L}_{\mathrm{fix}}^{\alpha'}$, $\mathcal{L}_{\mathsf{A}}^{\alpha'}$ does not contain the predicate $P_{<\alpha'}$. The axioms and rules of $\mathsf{AH}_{\alpha'}$ are given as follows.[6]

Axioms

Ax0. Γ, A, for a true $\mathcal{L}_{\mathsf{PA}}^+$-literal (i.e., $(\mathcal{L}_0^+ \setminus \{U\})$-literal) A.

Ax1. $\Gamma, \phi(t), \neg\phi(s)$, for an atomic ϕ of the form $P_\xi(t)$ $(\xi < \alpha)$, $P_{<\xi'}(t)$ $(\xi' \leq \alpha)$ or $U(t)$ and closed terms s and t with $s^{\mathbb{N}} = t^{\mathbb{N}}$.

Ax2. $\Gamma, \neg P_{<\xi}(t)$, for an ordinal $\xi \leq \alpha$ and a closed term t with $|(t)_1| \not< \xi$.

Ax3. $\Gamma, \neg P_\alpha^0(t)$, for any closed term t.

Logical rules and Cut the same as $\mathsf{H}_{\alpha'}$ (i.e., those of ω-arithmetic).

Fixed-point rules for P_ξ and $P_{<\xi'}$ ($\xi < \alpha$, $\xi' \leq \alpha$) the same as $\mathsf{H}_{\alpha'}$ (i.e., the Fixed-point rules III and IV in [43, p.63] for $\xi < \alpha$ and $\xi' \leq \alpha$)

Fixed-point rules for P_α^η for closed terms s and t with $|t| = \alpha$,

$$\frac{\Gamma, \mathcal{A}(P_\alpha^\eta, P_{<\alpha}, s, t)}{\Gamma, P_\alpha^{\eta+1}(s)} \qquad \frac{\Gamma, \neg\mathcal{A}(P_\alpha^\eta, P_{<\alpha}, s, t)}{\Gamma, \neg P_\alpha^{\eta+1}(s)}$$

Limit rules for P_α^λ for a limit λ and a closed term s,

$$\frac{\Gamma, P_\alpha^\eta(s) \quad (\text{for some } \eta < \lambda)}{\Gamma, P_\alpha^\lambda(s)} \qquad \frac{\Gamma, \neg P_\alpha^\eta(s) \quad (\text{for all } \eta < \lambda)}{\Gamma, \neg P_\alpha^\lambda(s)}$$

[6] Since we have assumed that $\langle\cdot,\cdot\rangle$ is bijective, we can drop the conditions like $t \in \mathrm{Pair}$ from the original formulation in [43].

For each $\mathcal{L}_{\mathsf{A}}^{\alpha'}$-formula ϕ, we define its AH-rank, $rk_{\mathsf{A}}(\phi)$, as follows: If ϕ is a literal of the form $P_\alpha^\eta t$ or $\neg P_\alpha^\eta t$ then $rk_{\mathsf{A}}(\phi) = \omega \cdot \eta$; if ϕ is a literal of any other form, $rk_{\mathsf{A}}(\phi) = 0$; $rk_{\mathsf{A}}(\phi)$ for a compositional formula ϕ is inductively defined in the standard manner; e.g., $rk_{\mathsf{A}}(\psi_0 \wedge \psi_1) := \max\{rk_{\mathsf{A}}(\psi_0), rk_{\mathsf{A}}(\psi_1)\} + 1$. We write $\mathsf{AH}_{\alpha'} \vdash_\rho^\delta \Gamma$ (or $\mathsf{AH}_{\alpha'} \vdash_{<\rho}^{\leq\delta} \Gamma$) when Γ is derived in $\mathsf{AH}_{\alpha'}$ with the length δ ($< \delta$ resp.) and cut-rank (w.r.t. rk_{A}) ρ ($< \rho$, resp.).

Suppose $\mathsf{H}_{\alpha'} \vdash_\star^\delta \Gamma$ and Γ contains no $P_{<\alpha'}$; then its derivation contains no occurrence of $P_{<\alpha'}$. Given arbitrary η, by a straightforward modification of Cantini's asymmetric interpretation in [4], we obtain $\mathsf{AH}_{\alpha'} \vdash_{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}}^{\leq\max\{\varepsilon(\delta),\varepsilon(\eta)\}} \Gamma[\eta, \eta+\omega^\delta]$, where $\Gamma[\xi_0, \xi_1]$ is obtained from Γ by replacing each occurrence of $\neg P_\alpha$ and P_α respectively by $\neg P_\alpha^{\xi_0}$ and $P_\alpha^{\xi_1}$.[7] Let β be $\max\{\varepsilon(\delta), \varepsilon(\eta)\}$. Then, by predicative cut elimination, we can obtain $\mathsf{AH}_{\alpha'} \vdash_1^{<\varphi 0\beta 0} \Gamma[\eta, \eta+\omega^\delta]$. We omit the details of this embedding of $\mathsf{H}_{\alpha'}$ in $\mathsf{AH}_{\alpha'}$ via asymmetric interpretation, but we will carry out a very similar embedding in Lemma 5.4.7 and the reader may refer to it. We state the so-called *Persistency Lemma* in $\mathsf{AH}_{\alpha'}$ for later use:

$$\text{If } \xi \leq \xi' \leq \eta' \leq \eta \text{ and } \mathsf{AH}_{\alpha'} \vdash_\rho^\delta \Gamma[\xi', \eta'], \Delta, \text{ then } \mathsf{AH}_{\alpha'} \vdash_{<\max\{\varepsilon(\eta),\varepsilon(\rho)\}}^{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}} \Gamma[\xi, \eta], \Delta; \tag{5.2}$$

this property is indeed heavily used in the above embedding of $\mathsf{H}_{\alpha'}$ in $\mathsf{AH}_{\alpha'}$.[8]

When $\Gamma \subset \mathcal{L}_{\mathrm{fix}}^\alpha$ (and thus $\Gamma[\xi_0, \xi_1] \equiv \Gamma$ for any ξ_0 and ξ_1), we take $\eta = 0$ and obtain $\mathsf{AH}_{\alpha'} \vdash_1^{<\varphi 0\varepsilon(\delta)0} \Gamma$. We can assume without loss of generality that its derivation contains no rules for P_α^ξ and thus can be regarded as a derivation of $\mathsf{H}_\alpha \vdash^{<\varphi 0\varepsilon(\delta)0} \Gamma$ in which cut is only applied to literals.[9] Then, we eliminate all the cuts applied to the literals of the forms $P_{<\beta}$ and $\neg P_{<\beta}$ ($\beta \leq \alpha$) or arithmetical literals without changing the upperbound of the length of derivation[10]; we have

[7] In fact, we can primitive recursively compute, from given δ and η, the exact length and cut-rank of $\mathsf{AH}_{\alpha'} \vdash_{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}}^{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}} \Gamma[\eta, \eta+\omega^\delta]$: that is, we have $\zeta_0(\eta,\delta), \zeta_1(\eta,\delta) < \max\{\varepsilon(\eta), \varepsilon(\delta)\}$ such that $\mathsf{AH}_{\alpha'} \vdash_{\zeta_1(\eta,\delta)}^{\zeta_0(\eta,\delta)} \Gamma[\eta, \eta+\omega^\delta]$. For example, we can take $\zeta_0(\eta,\delta) = \zeta_1(\eta,\delta) = \omega \cdot (\eta + \omega^\delta)$.

[8] The length and cut-rank in (5.2) are both primitive recursively computed as well.

[9] In general, if $\mathsf{AH}_{\alpha'} \vdash_\rho^\delta P_\alpha^0 t, \Gamma$, then $\mathsf{AH}_{\alpha'} \vdash_\rho^\delta \Gamma$; for, an atomic of the form $P_\alpha^0 t$ cannot be critical in any derivation.

[10] As a matter of fact, we can assume that the above derivation of $\mathsf{AH}_{\alpha'} \vdash_1^{<\varphi 0\varepsilon(\delta)0} \Gamma$. contains no cut to literals of

thus obtained $\mathsf{H}_\alpha|\!\frac{<\varphi 0\varepsilon(\delta)0}{\star}\ \Gamma$. In summary,

$$\text{If } \mathsf{H}_{\alpha'}|\!\frac{\delta}{\star}\ \Gamma \text{ for } \Gamma \text{ without } P_{<\alpha'}, \text{then } \mathsf{AH}_{\alpha'}|\!\frac{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}}{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}}\ \Gamma[\eta,\eta+\omega^\delta] \tag{5.3}$$

$$\text{If } \mathsf{H}_{\alpha'}|\!\frac{\delta}{\star}\ \Gamma \text{ for } \Gamma \subset \mathcal{L}^\alpha_{\mathrm{fix}}, \text{then } \mathsf{H}_\alpha|\!\frac{\delta'}{\star}\Gamma \text{ for some } \delta' < \varphi 0\varepsilon(\delta)0. \tag{5.4}$$

In fact, δ' in (5.4) can be primitive recursively computed from δ (see fn.7). Let $\mathcal{G}$ be a function such that $\mathcal{G}(\delta)$ is such δ'. We write $\mathcal{G}_k(\delta)$ for $\mathcal{G}\circ\cdots\circ\mathcal{G}(\delta)$ (k-times) and stipulate that $\mathcal{G}_0(\delta)=\delta$. We can assume that $\mathcal{G}(\delta')<\mathcal{G}(\delta)$ for $\delta'<\delta$; note that $\mathcal{G}_k(\delta)<(\delta|k)$. By iterating the procedure (5.4) for finite times, we have

$$\text{If } \mathsf{H}_{\alpha+m}|\!\frac{\delta}{\star}\ \Gamma \text{ for } \Gamma\subset\mathcal{L}^{\alpha+n}_{\mathrm{fix}} \text{ and } n<m, \text{ then } \mathsf{H}_{\alpha+n}|\!\frac{\mathcal{G}_{m-n}(\delta)}{\star}\ \Gamma.$$

The proof of Main Lemma II of [43] is essentially the transfinite repetition of the procedure (5.4). What we will do for our corresponding lemma (Lemma 5.4.10 below) essentially amounts to transfinitely iterating the procedure of Lemma 5.4.8 below instead of (5.4): i.e., the procedure of transforming $\mathsf{H}_{\alpha'}|\!\frac{\delta}{\star}\ \neg\mathrm{Cons}_{\alpha'},\Gamma$ into $\mathsf{H}_\alpha|\!\frac{\mathcal{G}'(\delta)}{\star}\ \neg\mathrm{Cons}_\alpha,\Gamma$ ($\mathcal{G}'$ will be given later).

Lemma 5.4.5. Let $\rho>0$. For arbitrary η_0 and η_1, if (a) $\mathsf{AH}_{\alpha'}|\!\frac{\delta_0}{\rho}P^{\eta_0}_\alpha s,\Gamma$ and (b) $\mathsf{AH}_{\alpha'}|\!\frac{\delta_1}{\rho}P^{\eta_1}_\alpha\dot{\neg}s,\Gamma$, then $\mathsf{AH}_{\alpha'}|\!\frac{\delta_0\#\delta_1}{\rho}\ \Gamma$.

Proof. This lemma informally corresponds to the fact that the minimal Kripkean fixed-point is consistent; in $\mathsf{AH}_{\alpha'}$, P^η_α corresponds to the η-th stage of the Kripkean construction (of the α-th truth) from the empty set, since we start with $\neg P^0_\alpha t$ for all t.

We show the claim by induction on $\delta_0\#\delta_1$. The base case is easy, since an atomic of the form $P^\eta_\alpha t$ cannot be critical in any axiom. For the induction step, the cases where either $P^{\eta_0}_\alpha s$ or $P^{\eta_1}_\alpha\dot{\neg}s$ is not critical in the last inference follow from IH. Let us assume $P^{\eta_0}_\alpha s$ and $P^{\eta_1}_\alpha\dot{\neg}s$ are both critical

the forms $P_{<\beta}$ and $\neg P_{<\beta}$ ($\beta\le\alpha$) and thus this elimination procedure is in fact redundant here.

in the last inferences. When either of the last inferences is by the limit rule, the premises contain $\mathsf{AH}_{\alpha'} \vdash^{\delta_2}_{\rho} P^{\eta_2}_{\alpha}(\dot{\neg})s, \Gamma$ for $\delta_2 < \delta_0$ (or $< \delta_1$) and $\eta_2 < \eta_0$ (or $< \eta_1$); then, the claim follows from IH. Hence, we focus on the essential case in which the last inferences are both the fixed-point rule. Then, we can assume that the premises are

$$\text{(c) } \mathsf{AH}_{\alpha'} \vdash^{\delta'_0}_{\rho} \Gamma, \mathcal{A}(P^{\eta_0 - 1}_{\alpha}, P_{<\alpha}, s, t) \quad \text{and} \quad \text{(d) } \mathsf{AH}_{\alpha'} \vdash^{\delta'_1}_{\rho} \Gamma, \mathcal{A}(P^{\eta_1 - 1}_{\alpha}, P_{<\alpha}, \dot{\neg}s, t),$$

for some $|t| = \alpha$, $\delta'_0 < \delta_0$ and $\delta'_1 < \delta_1$. By the form of $\mathcal{A}$ (and inversion), we can assume that $s \in \mathrm{St}_t$; let $s = \ulcorner \sigma \urcorner$. The claim is shown by cases according to the form of σ.

Suppose no negation '$\neg$' is attached to σ. We illustrate two cases. First, assume σ is $\phi \wedge \psi$. By $\wedge$-inversion, $\forall$-inversion and $\vee$-exportation, it follows from (c) that

$$\mathsf{AH}_{\alpha'} \vdash^{\delta'_0}_{\rho} \Gamma,\ \neg\mathrm{St}_t(\ulcorner \phi \urcorner),\ \neg\mathrm{St}_t(\ulcorner \psi \urcorner),\ s \neq \ulcorner \phi \wedge \psi \urcorner,\ P^{\eta_0 - 1}_{\alpha}(\ulcorner \phi \urcorner) \wedge P^{\eta_0 - 1}_{\alpha}(\ulcorner \psi \urcorner);$$

check with the form of $\mathcal{A}$. Since $\neg\mathrm{St}_t(\ulcorner \phi \urcorner), \neg\mathrm{St}_t(\ulcorner \psi \urcorner), s \neq \ulcorner \phi \wedge \psi \urcorner$ are all false $\mathcal{L}^{+}_{\mathsf{PA}}$-literals, we have $\mathsf{AH}_{\alpha'} \vdash^{\delta'_0}_{\rho} \Gamma, P^{\eta_0 - 1}_{\alpha}(\ulcorner \phi \urcorner) \wedge P^{\eta_0 - 1}_{\alpha}(\ulcorner \psi \urcorner)$. Then, by $\wedge$-inversion,

$$\mathsf{AH}_{\alpha'} \vdash^{\delta'_0}_{\rho} \Gamma, P^{\eta_0 - 1}_{\alpha}(\ulcorner \phi \urcorner) \quad \text{and} \quad \mathsf{AH}_{\alpha'} \vdash^{\delta'_0}_{\rho} \Gamma, P^{\eta_0 - 1}_{\alpha}(\ulcorner \psi \urcorner). \tag{5.5}$$

It also follows in the same way from (d) that

$$\mathsf{AH}_{\alpha'} \vdash^{\delta'_1}_{\rho} \Gamma, P^{\eta_1 - 1}_{\alpha}(\ulcorner \neg\phi \urcorner) \vee P^{\eta_1 - 1}_{\alpha}(\ulcorner \neg\psi \urcorner). \tag{5.6}$$

Now, we will show by side induction on δ'_1 that, under the (main) induction hypothesis up to $< \delta_0 \# \delta_1$, (5.5) and (5.6) (for $\delta'_0 < \delta_1$ and $\delta'_1 < \delta_1$) implies that $\vdash^{\delta'_0 \# \delta'_1}_{\rho} \Gamma$. The base step in which (5.6) is obtained by axiom is trivial. For the induction steps, the essential case is that

$P_\alpha^{\eta_1-1}(\ulcorner\neg\phi\urcorner) \vee P_\alpha^{\eta_1-1}(\ulcorner\neg\psi\urcorner)$ is critical in the last inference; otherwise the claim follows from SIH. In this case, the premise is

$$\mathsf{AH}_{\alpha'}\vdash^{\delta_1''}_{\rho} \Gamma, P_\alpha^{\eta_1-1\ulcorner\neg\phi\urcorner} \quad \text{or} \quad \mathsf{AH}_{\alpha'}\vdash^{\delta_1''}_{\rho} \Gamma, P_\alpha^{\eta_1-1\ulcorner\neg\psi\urcorner},$$

where $\delta_1'' < \delta_1'$. Then our claim follows from IH and (5.5).

Second, assume $s = \ulcorner\sigma\urcorner = \dot{T}_b r$ for $b, r \in \mathrm{CT}$ with $b^\circ < \alpha$. By $\wedge$-inversion, $\forall$-inversion and $\vee$-exportation, it follows from (c) that

$$\mathsf{AH}_{\alpha'}\vdash^{\delta_0'}_{\rho} \Gamma,\ \neg\mathrm{CT}(b),\ \neg\mathrm{CT}(r),\ b^\circ \not< t,\ s \neq \dot{T}_b r,\ P_{<\alpha}(\langle r^\circ, b^\circ\rangle).$$

Thus we obtain $\mathsf{AH}_{\alpha'}\vdash^{\delta_0'}_{\rho} \Gamma, P_{<\alpha}(\langle r^\circ, b^\circ\rangle)$, since $\neg\mathrm{CT}(b)$, $\neg\mathrm{CT}(r)$, $b^\circ \not< t$ and $s \neq \dot{T}_b r$ are all false $\mathcal{L}_0^+$-literals. It also follows in the same way from (d) that $\mathsf{AH}_{\alpha'}\vdash^{\delta_1'}_{\rho} \Gamma, \neg P_{<\alpha}(\langle r^\circ, b^\circ\rangle)$. Finally, we obtain, by cut (with rank 0), that $\mathsf{AH}_{\alpha'}\vdash^{\delta_0\#\delta_1}_{\rho} \Gamma$. The cases where σ is of another form can be shown similarly; but note that the case where $s = \dot{T}_b r$ for $b^\circ = \alpha$ needs neither the side-induction step nor cut with rank 0 and the claim directly follows from IH.

Next, assume σ is of the form $\neg\sigma'$; then $\dot{\neg} s = \ulcorner\neg\neg\sigma'\urcorner$. In the same manner as above, we obtain from (d) that $\mathsf{AH}_{\alpha'}\vdash^{\delta_1'}_{\rho} P_\alpha^{\eta_1-1}(\ulcorner\sigma\urcorner), \Gamma$. By IH, the claim $\mathsf{AH}_{\alpha'}\vdash^{\delta_0\#\delta_1}_{\rho} \Gamma$ follows from this and (c); the proof is completed. □

Lemma 5.4.6. (1) If $\mathsf{AH}_{\alpha'}\vdash^{\delta}_{\rho} P_\alpha^{\eta_0} s \wedge P_\alpha^{\eta_1}\dot{\neg} s, \Gamma$, for $\rho > 0$, then $\mathsf{AH}_{\alpha'}\vdash^{\delta\cdot 2}_{\rho} \Gamma$; by $\wedge$-inversion and the last lemma.

(2) For any closed term t, if $\mathsf{AH}_{\alpha'}\vdash^{\delta}_{\rho} \exists x \in \mathrm{St}_t[P_\alpha^{\eta_0}(x) \wedge P_\alpha^{\eta_1}(\dot{\neg} x)], \Gamma$, for $\rho > 0$, then $\mathsf{AH}_{\alpha'}\vdash^{\delta\cdot 2}_{\rho} \Gamma$; by induction on δ.

Lemma 5.4.7. Let Γ contain no $P_{<\alpha'}$. Suppose $\mathsf{H}_{\alpha'}\vdash^{\delta}_{\star} \neg\mathrm{Cons}_{\alpha'}, \Gamma$. Then, for each η, we have $\mathsf{AH}_{\alpha'}\vdash^{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}}_{<\max\{\varepsilon(\delta),\varepsilon(\eta)\}} \neg\mathrm{Cons}_\alpha, \Gamma[\eta, \eta+\omega^{1+\delta}]$; in fact, the length and cut-rank can be computed to

be $\omega \cdot (\eta + \omega^{1+\delta}) \cdot 2$ (cf. fn.7 and fn.8).

Proof. Suppose $\mathsf{H}_{\alpha'} \vdash^{\delta}_{\star} \neg\mathrm{Cons}_{\alpha'}, \Gamma$. The claim is shown by induction on δ. The proof is essentially parallel to that of the embedding (5.3) except that $P_{<\alpha'}$ occurs in $\neg\mathrm{Cons}_{\alpha'}$. Thus, each case where Γ is critical in the last inference is dealt with in a parallel manner; we illustrate some typical cases below and the rest are similarly shown. For simplicity, we write $||\xi_0, \xi_1||$ for $\max\{\varepsilon(\xi_0), \varepsilon(\xi_1)\}$.

Suppose Γ is obtained from the premises $\mathsf{H}_{\alpha'} \vdash^{\delta_0}_{\star} \neg\mathrm{Cons}_{\alpha'}, \Gamma, P_\alpha t$ and $\mathsf{H}_{\alpha'} \vdash^{\delta_1}_{\star} \neg\mathrm{Cons}_{\alpha'}, \Gamma, \neg P_\alpha t$ for some $\delta_0, \delta_1 < \delta$ by cut. By IH, we have

$$\mathsf{AH}_{\alpha'} \vdash \frac{< ||\delta_0, \eta||}{< ||\delta_0, \eta||}\ \neg\mathrm{Cons}_\alpha, \Gamma[\eta,\ \eta + \omega^{1+\delta_0}], P_\alpha^{\eta+\omega^{1+\delta_0}} t$$
$$\mathsf{AH}_{\alpha'} \vdash \frac{< ||\delta_1, \eta + \omega^{1+\delta_0}||}{< ||\delta_1, \eta + \omega^{1+\delta_0}||}\ \neg\mathrm{Cons}_\alpha, \Gamma[\eta + \omega^{1+\delta_0},\ \eta + \omega^{1+\delta_0} + \omega^{1+\delta_1}], \neg P_\alpha^{\eta+\omega^{1+\delta_0}} t.$$

Since $\varepsilon(\eta + \omega^{1+\delta}), ||\delta_0, \eta||, ||\delta_1, \eta + \omega^{1+\delta_0}|| \leq ||\delta, \eta||$, we obtain

$$\mathsf{AH}_{\alpha'} \vdash \frac{< ||\delta, \eta||}{< ||\delta, \eta||}\ \neg\mathrm{Cons}_\alpha, \Gamma[\eta, \eta + \omega^{1+\delta}], P_\alpha^{\eta+\omega^{1+\delta_0}} t$$
$$\mathsf{AH}_{\alpha'} \vdash \frac{< ||\delta, \eta||}{< ||\delta, \eta||}\ \neg\mathrm{Cons}_\alpha, \Gamma[\eta, \eta + \omega^{1+\delta}], \neg P_\alpha^{\eta+\omega^{1+\delta_0}} t,$$

by persistency (5.2). Since $\omega \cdot (\eta + \omega^{1+\delta_0}) < ||\delta, \eta||$, the claim follows by cut.

Suppose Γ is obtained from the premise $\mathsf{H}_{\alpha'} \vdash^{\delta'}_{\star} \neg\mathrm{Cons}_{\alpha'}, \Gamma', \mathcal{A}(P_\alpha, P_{<\alpha}, t, s)$, for $\delta' < \delta$, $|s| = \alpha$ and $\Gamma = \Gamma' \cup \{P_\alpha t\}$, by Fixed-point rule. Then, by applying (5.2) to IH in the same manner as above, we obtain

$$\mathsf{AH}_{\alpha'} \vdash \frac{< ||\delta, \eta||}{< ||\delta, \eta||}\ \neg\mathrm{Cons}_\alpha, \Gamma[\eta, \eta + \omega^{1+\delta}], \mathcal{A}(P_\alpha^{\eta+\omega^{1+\delta'}}, P_{<\alpha}, t, s).$$

By the same Fixed-point rule, we obtain

$$\mathsf{AH}_{\alpha'}\vdash^{<||\delta,\eta||}_{<||\delta,\eta||}\neg\mathrm{Cons}_\alpha, \Gamma[\eta, \eta+\omega^{1+\delta}], P_\alpha^{\eta+\omega^{1+\delta'}+1}(t);$$

the claim follows by (5.2), since $\varepsilon(\eta+\omega^{1+\delta'}+1) \leq ||\delta,\eta||$.

Now we turn to the essential case in which $\neg\mathrm{Cons}_{\alpha'}$ is critical in the last inference. Then the premise is: for some $\delta' < \delta$ and closed term t,

$$\mathsf{H}_{\alpha'}\vdash^{\delta'}_{\star} t < \alpha' \wedge \exists x \in \mathrm{St}_t[P_{<\alpha'}(\langle x,t\rangle) \wedge P_{<\alpha'}(\langle \dot{\neg} x,t\rangle)],\ \Gamma.$$

If $|t| \not< \alpha'$, the claim trivially follows. Let $|t| < \alpha'$. By $\wedge$-inversion, we have

$$\mathsf{H}_{\alpha'}\vdash^{\delta'}_{\star} (\exists x \in \mathrm{St}_t)[P_{<\alpha'}\big(\langle x,t\rangle\big) \wedge P_{<\alpha'}\big(\langle \dot{\neg} x,t\rangle\big)],\ \Gamma.$$

If $|t| < \alpha$, then it follows from Lemma 5.4.4 that

$$\mathsf{H}_{\alpha'}\vdash^{2+\delta'}_{\star} (\exists x \in \mathrm{St}_t)[P^\eta_{<\alpha}\big(\langle x,t\rangle\big) \wedge P^\eta_{<\alpha}\big(\langle \dot{\neg} x,t\rangle\big)],\ \Gamma.$$

Hence, we have $\mathsf{AH}_{\alpha'}\vdash^{<||1+\delta,\eta||}_{<||1+\delta,\eta||}\exists x \in \mathrm{St}_t[P^\eta_{<\alpha}(\langle x,t\rangle) \wedge P^\eta_{<\alpha}(\langle \dot{\neg} x,t\rangle)], \Gamma[\eta, \eta+\omega^{1+\delta}]$ by the procedure (5.3); we obtain the claim by Ax0 and logical rules. Finally, assume $|t| = \alpha$. By Lemma 5.4.3, we have $\mathsf{H}_{\alpha'}\vdash^{1+\delta'}_{\star} (\exists x \in \mathrm{St}_t)[P_\alpha(x) \wedge P_\alpha(\dot{\neg} x)], \Gamma$; therefore, since $1+\delta' \leq \delta$, we have

$$\mathsf{AH}_{\alpha'}\vdash^{\beta_0}_{\beta_1} (\exists x \in \mathrm{St}_t)[P_\alpha^{\eta+\omega^{1+\delta}}(x) \wedge P_\alpha^{\eta+\omega^{1+\delta}}(\dot{\neg} x)], \Gamma[\eta, \eta+\omega^{1+\delta}],$$

for some $\beta_0, \beta_1 < ||\delta,\eta||$, by the procedure (5.3). By the last lemma, we obtain $\mathsf{AH}_{\alpha'}\vdash^{\beta_0\cdot 2}_{\beta_1} \Gamma[\eta, \eta+\omega^{1+\delta}]$; the claim follows from this since $\beta_0 \cdot 2 < ||\delta,\eta||$. $\square$

Lemma 5.4.8. Let $\Gamma \subset \mathcal{L}^{\alpha}_{\mathrm{fix}}$. If $\mathsf{H}_{\alpha'}\vdash^{\delta}_{\star} \neg\mathrm{Cons}_{\alpha'}, \Gamma$, then $\mathsf{H}_{\alpha}\vdash^{<\varphi 0\varepsilon(\delta)0}_{\star} \neg\mathrm{Cons}_{\alpha}, \Gamma$.

Proof. Suppose $\mathsf{H}_{\alpha'}\vdash^{\delta}_{\star}\neg\mathrm{Cons}_{\alpha'}, \Gamma$. Taking $\eta = 0$ in the last lemma, we obtain $\mathsf{AH}_{\alpha'}\vdash^{<\varepsilon(\delta)}_{<\varepsilon(\delta)}\neg\mathrm{Cons}_{\alpha}, \Gamma$. By predicative cut elimination, $\mathsf{AH}_{\alpha'}\vdash^{<\varphi 0\varepsilon(\delta)0}_{1}\neg\mathrm{Cons}_{\alpha}, \Gamma$. Since this sequent is of $\mathcal{L}^{\alpha}_{\mathrm{fix}}$, we can replace $\mathsf{AH}_{\alpha'}\vdash^{<\varphi 0\varepsilon(\delta)0}_{1}$ by $\mathsf{H}_{\alpha}\vdash^{<\varphi 0\varepsilon(\delta)0}_{\star}$. □

It is observed from this proof (and Lemma 5.4.7) that we can primitive recursively compute the length of $\mathsf{H}_{\alpha}\vdash^{<\varphi 0\varepsilon(\delta)0}_{\star}\neg\mathrm{Cons}_{\alpha}, \Gamma$, from given δ; in fact, we can take $\varphi 0(\omega^{1+\delta}\cdot 2)(\omega^{1+\delta}\cdot 2)$ as such for instance. Let $\mathcal{G}'(\delta)$ is such a length: i.e., $\mathcal{G}'(\delta) < \varphi 0\varepsilon(\delta)0$ and $\mathsf{H}_{\alpha}\vdash^{\mathcal{G}'(\delta)}_{\star}\neg\mathrm{Cons}_{\alpha}, \Gamma$, for $\Gamma \subset \mathcal{L}^{\alpha}_{\mathrm{fix}}$, when $\mathsf{H}_{\alpha'}\vdash^{\delta}_{\star}\neg\mathrm{Cons}_{\alpha'}, \Gamma$. For each $k \in \mathbb{N}$, $\mathcal{G}'_k(\delta)$ is defined in the analogous way to $\mathcal{G}_k(\delta)$; thus $\mathcal{G}'_k(\delta) < (\delta|k)$.

Corollary 5.4.9. Let $m, k \in \mathbb{N}$ and $\Gamma \subset \mathcal{L}^{\alpha+m}_{\mathrm{fix}}$. If $\mathsf{H}_{\alpha+m+k}\vdash^{\delta}_{\star} \neg\mathrm{Cons}_{\alpha+m+k}, \Gamma$, then we have $\mathsf{H}_{\alpha+m}\vdash^{\mathcal{G}'_k(\delta)}_{\star} \neg\mathrm{Cons}_{\alpha+m}, \Gamma$.

Lemma 5.4.10. Suppose $\mathsf{H}_{\alpha+\omega^{1+\rho}}\vdash^{\delta}_{\star} \neg\mathrm{Cons}_{\alpha+\omega^{1+\rho}}, \Gamma$ for $\Gamma \subset \mathcal{L}^{\alpha+\xi}_{\mathrm{fix}}$ and $\xi < \omega^{1+\rho}$. Then we have $\mathsf{H}_{\alpha+\xi}\vdash^{\varphi 1\rho(2+\delta)}_{\star} \neg\mathrm{Cons}_{\alpha+\xi}, \Gamma$.

Proof. This corresponds to Main Lemma II of [43]. We prove the claim by induction on ρ with side induction on δ. The proof is divided into three cases: $\rho = 0$; ρ is successor; ρ is limit. We only illustrate the successor case; the other cases can be shown in the same pattern; but note that we use the last corollary for the base case ($\rho = 0$) instead of IH.

Let $\rho = \rho' + 1$. Since $\xi < \omega^{1+\rho'+1}$, there is a natural number n such that $\xi = \omega^{1+\rho'}\cdot n + \eta$ for some $\eta < \omega^{1+\rho'}$. If $\delta = 0$ then Γ is critical and the claim trivially obtains. Let us assume $\delta > 0$. If Γ is critical in the last inference and it is not cut, the claim follows from SIH. The crucial cases are where $\neg\mathrm{Cons}_{\alpha+\omega^{1+\rho}}$ is critical and where the last inference is cut.

Suppose the first case. By $\wedge$-inversion, $\mathsf{H}_{\alpha+\omega^{1+\rho}}\vdash^{\delta'}_{\star} t < \alpha + \omega^{1+\rho}, \Gamma$ and

$$\mathsf{H}_{\alpha+\omega^{1+\rho}}\vdash^{\delta'}_{\star} \exists x \in \mathrm{St}_t[P_{<\alpha+\omega^{1+\rho}}(\langle x, t\rangle) \wedge P_{<\alpha+\omega^{1+\rho}}(\langle \dot{\neg} x, t\rangle)], \Gamma,$$

for some t and $\delta' < \delta$. The case where $|t| \not< \alpha + \omega^{1+\rho}$ is trivially obtained and we assume $|t| < \alpha+\omega^{1+\rho}$. The case in which $|t| < \alpha+\xi$ follows from Lemma 5.4.4. Let $\alpha+\xi \leq |t| < \alpha+\omega^{1+\rho'}\cdot m$ for some natural number $m > n$. It follows from Lemma 5.4.4 that

$$\mathsf{H}_{\alpha+\omega^{1+\rho}} \vdash \frac{2+\delta'}{\star}\ \exists x \in \mathrm{St}_t[P_{<\alpha+\omega^{1+\rho'}\cdot m}(\langle x, t\rangle) \wedge P_{<\alpha+\omega^{1+\rho'}\cdot m}(\langle \dot{\neg} x, t\rangle)],\ \Gamma. \tag{5.7}$$

Since this sequent is of $\mathcal{L}_{\mathrm{fix}}^{\alpha+\omega^{1+\rho'}\cdot m}$, we can apply Main Lemma II of [43] to (5.7) and obtain

$$\mathsf{H}_{\alpha+\omega^{1+\rho'}\cdot m} \vdash \frac{\varphi 1\rho(2+\delta')}{\star}\ \exists x \in \mathrm{St}_t[P_{<\alpha+\omega^{1+\rho'}\cdot m}(\langle x, t\rangle) \wedge P_{<\alpha+\omega^{1+\rho'}\cdot m}(\langle \dot{\neg} x, t\rangle)],\ \Gamma;$$

thus we have $\mathsf{H}_{\alpha+\omega^{1+\rho'}\cdot m} \vdash \frac{\varphi 1\rho(2+\delta')+2}{\star}\ \neg\mathrm{Cons}_{\alpha+\omega^{1+\rho'}\cdot m}, \Gamma$. Since we have assumed that $\xi < \omega^{1+\rho'} \cdot (n+1) \leq \omega^{1+\rho'} \cdot m$, we obtain by applying IH for $(m-n)$-times (cf. [43]) that

$$\mathsf{H}_{\alpha+\xi} \vdash \frac{\varphi 1\rho'\big(\cdots\big(\varphi 1\rho'\big(\varphi 1\rho(2+\delta')+2\big)\big)\cdots\big)}{\star}\ \neg\mathrm{Cons}_{\alpha+\xi}, \Gamma,$$

where $\varphi 1\rho'(\cdots(\varphi 1\rho'(\varphi 1\rho(2+\delta')+2))\cdots)$ is the result of operating $\varphi 1\rho'(\cdot)$ for $(m-n)$-times. Since $\rho' < \rho$ and $\varphi 1\rho(2+\delta')+2 < \varphi 1\rho(2+\delta)$, it is $< \varphi 1\rho(2+\delta)$; we are done.

The case in which the last inference is cut remains to be shown; but this can be shown in a completely parallel manner to the one in the cited lemma of [43] (use SIH first, apply cut, and then use IH finitely many times). □

Corollary 5.4.11. Let $\Gamma \subset \mathcal{L}_0^+$ and $\alpha = \omega^{1+\alpha_n} + \cdots + \omega^{1+\alpha_1} + m$. If $\mathsf{H}_\alpha \vdash \frac{\delta}{\star} \neg\mathrm{Cons}_\alpha, \Gamma$ then $\mathsf{H}_0 \vdash \frac{<\varphi 1\alpha_n(\cdots(\varphi 1\alpha_1((\alpha|m)))\cdots)}{\star}\ \neg\mathrm{Cons}_0, \Gamma$.

Lemma 5.4.12. If $\mathsf{H}_\alpha \vdash^\delta \neg\mathrm{Cons}_0, \Gamma$, then $\mathsf{H}_\alpha \vdash^\delta \Gamma$; by induction on δ (cf. $\neg\mathrm{Cons}_0$ is simply false).

Theorem 5.4.13. $|\mathsf{KF}_\alpha + \mathrm{Cons}_\alpha| = |\mathsf{KF}_\alpha| = |\mathsf{KF}_\alpha + \mathrm{Comp}_\alpha|$. It also follows, in the same way as Theorem 2 of [78], that $|\mathsf{Aut}(\mathsf{KF}+\mathrm{Cons})| = |\mathsf{Aut}(\mathsf{KF}+\mathrm{Comp})| = \varphi 200$.

5.5 Positive Uniform T-biconditionals

Recall that the purely disquotationalist system PUTB comprises the schema of *positive uniform T-biconditionals*, PUTB: namely, $\phi(\vec{x})$ is true iff $\phi(\vec{x})$, for each formula $\phi(\vec{x})$ in which the truth predicate occurs only positively. In the present section, we consider the iteration and autonomous progression of this schema.

Definition 5.5.1. Let $\prec$ be a p.r. ordering and $T^{\prec}_{\prec z}(x,y)$ denote $y \prec z \wedge T^{\prec}_{y}(x)$. The system $\mathsf{PUTB}_{\prec}$ over $\mathcal{L}^{\prec}$ is defined as $\mathsf{PA}(U)$ plus full-induction for $\mathcal{L}^{\prec}$, $\mathrm{TI}_{\mathcal{L}^{\prec}}(\prec)$ and the axiom schema $\mathrm{PUTB}_{\prec}$: for each R-positive formula $\mathcal{A}(R,Q,\vec{x},y) \in \mathcal{L}_0(R,Q)$,

$$\forall y \in \mathrm{fd}(\prec)\forall \vec{x}\big[\mathcal{A}(T^{\prec}_{y}, T^{\prec}_{\prec y}, \vec{x}, y) \leftrightarrow T^{\prec}_{y}\big(\ulcorner \mathcal{A}(T^{\prec}_{\dot{y}}, T^{\prec}_{\prec \dot{y}}, \vec{\dot{x}}, \dot{y})\urcorner\big)\big].$$

The system $\mathsf{Aut}(\mathsf{PUTB})$ denotes $\mathsf{PA}(U)$ plus full-induction for $\mathcal{L}$ and the following rules:

$$\frac{\mathrm{TI}(\prec; U)}{\phi} \qquad \text{and} \qquad \frac{\mathrm{TI}(\prec; U)}{\mathrm{TI}(\prec; \psi)},$$

for each p.r. linear ordering $\prec$, any instance ϕ of $\mathrm{PUTB}_{\prec}$ and arbitrary $\mathcal{L}$-formula ψ.

Lemma 5.5.2. There exists a syntactical embedding of $\mathsf{PUTB}_{\prec}$ in $\mathsf{KF}_{\prec}$.

Proof. It is known that $\mathsf{KF} \vdash \mathrm{PUTB}$ and its proof is a straightforward meta-induction; see [5]. However, the same proof doesn't work for iterated truth and we will construct a suitable translation of $\mathsf{PUTB}_{\prec}$ in $\mathsf{KF}_{\prec}$. By the recursion theorem (or diagonalization), we define a binary function f such that

$$f(x,y) = \begin{cases} x & \text{if } x \in \mathrm{AtFml}_0 \\ \dot{T}^{\prec}_{\dot{y}}\ \mathrm{nm}\big(\ulcorner T^{\prec}_{z} f(w,z)\urcorner\big) & \text{if } x = \dot{T}^{\prec}_{\dot{z}} w \text{ for } z,w \in \mathrm{Tm} \\ \dot{\neg} f(z,y) \quad \big(f(z,y) \dot{\wedge} f(w,y), \text{ resp.}\big) & \text{if } x = \dot{\neg} z \text{ (or } x = z \dot{\wedge} w) \\ \dot{\forall} w.f(z,y) & \text{if } x = \dot{\forall} w.z; \end{cases}$$

otherwise $f(x,y)=0$; recall that $\ulcorner T^{\prec}_z f(w,z)\urcorner$ means $\mathrm{sb}(\ulcorner T^{\prec}(f(\cdot,\cdot),\cdot)\urcorner; w, z)$. Now, let $\mathcal{T}$ be a syntactical translation which preserves $\mathcal{L}_0$-atomics, translates $T^{\prec}_z x$ by $T^{\prec}_z f(x,z)$ and commutes with the logical connectives and quantifiers. We will show $\mathcal{T}$ is an interpretation of $\mathsf{PUTB}_{\prec}$ in $\mathsf{KF}_{\prec}$ by meta-induction on $\mathcal{A}$. First, we trivially have $\mathcal{T}\big(T^{\prec}_z(\ulcorner\phi\urcorner)\big) \leftrightarrow \mathcal{T}(\phi)$ for all $\mathcal{L}_0$-atomics ϕ. It is also easily observed that $\mathcal{T}\big(T^{\prec}_z(\ulcorner T^{\prec}_{\dot{z}} t(\vec{\dot{x}})\urcorner)\big)$ is equivalent to $\mathcal{T}(T^{\prec}_z t(\vec{x}))$ in $\mathsf{KF}_{\prec}$. Next, we show that $\mathcal{T}\big(T^{\prec}_{\prec z}(s(\vec{x}),t(\vec{x}))\big)$ and $\mathcal{T}\big(\neg T^{\prec}_{\prec z}(s(\vec{x}),t(\vec{x}))\big)$ are respectively equivalent to $\mathcal{T}\big(T^{\prec}_z(\ulcorner T^{\prec}_{\prec\dot{z}}(s(\vec{\dot{x}}),t(\vec{\dot{x}}))\urcorner)\big)$ and $\mathcal{T}\big(F^{\prec}_z(\ulcorner T^{\prec}_{\prec\dot{z}}(s(\vec{\dot{x}}),t(\vec{\dot{x}}))\urcorner)\big)$. We only demonstrate the latter:

$$\begin{aligned}
\mathcal{T}\big(F^{\prec}_z\big(\ulcorner t(\vec{\dot{x}}) \prec \dot{z} \wedge T^{\prec}_{t(\vec{\dot{x}})} s(\vec{\dot{x}})\urcorner\big)\big) &\leftrightarrow F^{\prec}_z\big(\ulcorner t(\vec{\dot{x}}) \prec \dot{z}\urcorner \mathbin{\dot{\wedge}} \dot{T}^{\prec}_{\dot{z}}\mathrm{nm}\big(\ulcorner \dot{T}^{\prec}_{t(\vec{\dot{x}})} f(s(\vec{\dot{x}}),t(\vec{\dot{x}}))\urcorner\big)\big)\\
&\leftrightarrow t(\vec{x}) \not\prec z \vee F^{\prec}_z\big(\dot{T}^{\prec}_{\dot{z}}\mathrm{nm}\big(\ulcorner \dot{T}^{\prec}_{t(\vec{\dot{x}})} f(s(\vec{\dot{x}}),t(\vec{\dot{x}}))\urcorner\big)\big)\\
&\leftrightarrow t(\vec{x}) \not\prec z \vee \neg T^{\prec}_{t(\vec{x})} f(s(\vec{x}),t(\vec{x})) \;\equiv \mathcal{T}\big(\neg T^{\prec}_{\prec z}(s(\vec{x}),t(\vec{x}))\big);
\end{aligned}$$

it is crucial here that $f(\dot{T}^{\prec}_w u, z)$ is always in $\mathrm{St}^{\prec}_z$ for any $u, w \in \mathrm{CT}$. We have thus established the base step. Finally, by observing that $f\big(\ulcorner\mathcal{A}(T^{\prec}_{\dot{y}}, T^{\prec}_{\prec\dot{y}}, \vec{\dot{x}}, \dot{y})\urcorner, y\big) \in \mathrm{St}^{\prec}_y$ for each $y \in \mathrm{fd}(\prec)$ and $\mathcal{A}$, the induction steps are straightforwardly obtained. □

Corollary 5.5.3. There exists a syntactical embedding of $\mathsf{PUTB}_{\prec} + \mathrm{Cons}_{\prec}$ ($\mathsf{PUTB}_{\prec} + \mathrm{Comp}_{\prec}$) in $\mathsf{KF}_{\prec} + \mathrm{Cons}_{\prec}$ ($\mathsf{KF}_{\prec} + \mathrm{Comp}_{\prec}$, resp.).

Proof. The above interpretation $\mathcal{T}$ works as well. Observe that, for any $z \in \mathrm{fd}(\prec)$ and $x \in \mathrm{St}^{\prec}_z$, $\mathcal{T}(T^{\prec}_z x \wedge F^{\prec}_z x)$ is equivalent to $T^{\prec}_z f(x,z) \wedge T^{\prec}_z \dot{\neg} f(x,z)$; the first claim follows from this. The second claim is shown similarly. □

Lemma 5.5.4. There is a syntactical embedding of $\widehat{\mathsf{ID}}_{\prec}$ in $\mathsf{PUTB}_{\prec}$ (and thus in $\mathsf{KF}_{\prec}$ as well).

Proof. Given an inductive operator $\mathcal{A}(R,Q,x,y)$, set $\mathcal{A}'(x,y,z) \in \mathcal{L}^{\prec}$ be:

$$\mathcal{A}\big(\lambda u.T_y^{\prec} z(u,y),\ \lambda w.[(w)_1 \prec y \wedge T^{\prec}_{(w)_1} z((w)_0,(w)_1)],\ x,\ y\big).$$

Recall that $z(u,v)$ is the result of respectively substituting u-th and v-th numerals for the first and second free variables in (a code of formula) z; we assume here that x comes earlier than y in the fixed variable enumeration. Then, by parametrized diagonalization (for z), there is a formula $\psi_{\mathcal{A}}(x,y)$ such that

$$\psi_{\mathcal{A}}(x,y) \ \leftrightarrow\ \mathcal{A}\big(\lambda u.T_y^{\prec\ulcorner}\psi_{\mathcal{A}}(\dot{u},\dot{y})^{\urcorner},\ \lambda w.T^{\prec}_{\prec y}\big(\ulcorner\psi((\dot{w})_0,(\dot{w})_1)^{\urcorner},(w)_1\big),\ x,\ y\big);$$

note that $(w)_1 \prec y \wedge T^{\prec}_{(w)_1}{}^{\ulcorner}\psi^{\urcorner}((w)_0,(w)_1) \equiv T^{\prec}_{\prec y}\big(\ulcorner\psi((\dot{w})_0,(\dot{w})_1)^{\urcorner},(w)_1\big)$. By the standard construction of $\psi_{\mathcal{A}}$, we can assume that $\psi_{\mathcal{A}}$ is of the form $\mathcal{B}(T_y^{\prec}, T^{\prec}_{\prec y}, x, y)$ for some R-positive $\mathcal{B} \in \mathcal{L}_0(R,Q)$ and thus $\mathsf{PUTB}_{\prec} \vdash \forall x \forall y\big[T_y^{\prec\ulcorner}\psi_{\mathcal{A}}(\dot{x},\dot{y})^{\urcorner} \leftrightarrow \psi_{\mathcal{A}}(x,y)\big]$. Then, the embedding * of $\widehat{\mathsf{ID}}_{\prec}$ in $\mathsf{PUTB}_{\prec}$ is obtained by $(P^{\mathcal{A}}x)^* \mapsto T_y^{\prec\ulcorner}\psi_{\mathcal{A}}((\dot{x})_0,(\dot{x})_1)^{\urcorner}$, since $(P^{\mathcal{A}}_{\prec y}w)^* \equiv T^{\prec}_{\prec y}\big(\ulcorner\psi((\dot{w})_0,(\dot{w})_1)^{\urcorner},(w)_1\big)$. □

Theorem 5.5.5. $|\mathsf{PUTB}_\alpha| = |\mathsf{PUTB}_\alpha + \mathrm{Cons}_\alpha| = |\mathsf{PUTB}_\alpha + \mathrm{Comp}_\alpha| = |\mathsf{KF}_\alpha|$. We also have $|\mathsf{Aut}(\mathsf{PUTB})| = |\mathsf{Aut}(\mathsf{PUTB} + \mathrm{Cons})| = |\mathsf{Aut}(\mathsf{PUTB} + \mathrm{Comp})| = \varphi 200$.

5.6 Kripke-Feferman Truth with Weak Kleene Logic

In the present section, we consider the iteration and autonomous progression of WKF-truth. I remind the reader that the truth tables for the logical connectives in weak Kleene logic are given as follows:

$\neg$	
T	F
U	U
F	T

$\wedge$	T	U	F
T	T	U	F
U	U	U	U
F	F	U	F

The evaluation of the universal quantifier is: $\forall x\phi(x)$ is true iff $\phi(x)$ is true for all x; and $\forall x\phi(x)$ is false iff $\phi(x)$ is false for some x and $\phi(x)$ is either true or false for all x. The other logical primitives are defined in the usual manner in terms of '$\neg$', '$\wedge$' and '$\forall$'.

Definition 5.6.1. Given a p.r. ordering $\prec$, the system $\mathsf{WKF}_\prec$ over $\mathcal{L}^\prec$ consists of $\mathsf{PA}(U)$, full induction for $\mathcal{L}^\prec$, $\mathrm{TI}_{\mathcal{L}^\prec}(\prec)$, $\mathbf{K1}_\prec$–$\mathbf{K4}_\prec$, and: (instead of **K5**–**K6**)

$\mathbf{W5}_\prec$ $\forall a{\in}\mathrm{fd}(\prec)\forall x, y \in \mathrm{St}_a^\prec$

$\big[\big(T_a^\prec(x\dot{\wedge}y) \leftrightarrow (T_a^\prec x \wedge T_a^\prec y)\big)$

$\wedge\big[F_a^\prec(x\dot{\wedge}y) \leftrightarrow \big((F_a^\prec x \wedge F_a^\prec y) \vee (T_a^\prec x \wedge F_a^\prec y) \vee (F_a^\prec x \wedge T_a^\prec y)\big)\big]\big]$.

$\mathbf{W6}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall z\forall x$

$\big[\forall z.x{\in}\mathrm{St}_a^\prec \rightarrow \big[\big(T_a^\prec(\dot{\forall}z.x) \leftrightarrow \forall y T_a^\prec x(y)\big)$

$\wedge\big(F_a^\prec(\dot{\forall}z.x) \leftrightarrow \big(\exists y F_a^\prec x(y) \wedge \forall y(T_a^\prec x(y) \vee F_a^\prec x(y))\big)\big)\big]\big]$.

The system $\mathsf{Aut}(\mathsf{WKF})$ is $\mathsf{PA}(U)$ plus full-induction for $\mathcal{L}$, and the following rules:

$$\frac{\mathrm{TI}(\prec; U)}{\mathbf{K1}_\prec \wedge \cdots \wedge \mathbf{K4}_\prec \wedge \mathbf{W5}_\prec \wedge \mathbf{W6}_\prec} \qquad \text{and} \qquad \frac{\mathrm{TI}(\prec; U)}{\mathrm{TI}(\prec; \psi)},$$

for each p.r. linear ordering $\prec$ and arbitrary $\mathcal{L}$-formula ψ.

$\mathsf{WKF}_\prec$ is embeddable in $\widehat{\mathsf{ID}}_\prec$ in the same way as in the case of $\mathsf{KF}_\prec$; for, $\mathsf{WKF}_\prec$ can be modelled by iterated Kripkean fixed-points with weak Kleene schema (cf. [50]). Hence, we can regard $\mathsf{WKF}_\prec \subset \widehat{\mathsf{ID}}_\prec$; therefore $|\mathsf{WKF}_\prec| \leq |\mathsf{KF}_\prec|$ and $|\mathsf{Aut}(\mathsf{WKF})| \leq |\mathsf{Aut}(\mathsf{KF})|$. In the present section, we give a well-ordering proof for WKF_α which yields $|\mathsf{WKF}_\alpha| \geq |\mathsf{KF}_\alpha|$. For readability, as in §4, we

identify ordinals with their representations in the fixed notation system and drop the superscript $\lhd \restriction_\alpha$ whenever there is no worry of confusion.

The well-ordering proof for WKF_α goes almost parallel to the one for KF_α (SRT_α) in [43]. The point is that KF_α and WKF_α behave in the same way on determinate sentences. However, we still need to make slight modifications. In what follows, instead of giving a well-ordering proof specifically for WKF_α, we will show a more general theorem.

Theorem 5.6.2. Given an ordinal α such that

$$\alpha = \omega^{1+\alpha_0} + \cdots + \omega^{1+\alpha_n} + m,$$

for ordinals $\alpha_0 \geq \ldots \geq \alpha_n$ and $m < \omega$, let Q_α be a theory over $\mathcal{L}^{\lhd\restriction_\alpha}$ which extends $\mathsf{PA}(U)$ and contains full-induction for $\mathcal{L}^{\lhd\restriction_\alpha}$ and $\mathrm{TI}_{\mathcal{L}^{\lhd\restriction_\alpha}}(\alpha)$. We write $D^+_\beta(x)$ for $\mathrm{St}_\beta(x) \wedge [F_\beta x \leftrightarrow \neg T_\beta x]$. Suppose Q_α derives all the following:

(i) The axioms $\mathbf{K1}_\alpha$ and $\mathbf{K2}_\alpha$;

(ii) $\forall\beta < \alpha \forall x \forall y \big[(x \in D^+_\beta \wedge y \in D^+_\beta) \rightarrow (\dot{\neg} x \in D^+_\beta \wedge x \dot{\wedge} y \in D^+_\beta)\big]$;

(iii) $\forall\beta < \alpha \forall x \forall z \big[\big(\dot{\forall} z.x \in \mathrm{St}_\beta \wedge \forall y\big(x(y) \in D^+_\beta\big)\big) \rightarrow (\dot{\forall} z.x \in D^+_\beta)\big]$;

(iv) $\forall\beta < \alpha \forall x \big[\dot{\neg} x \in D^+_\beta \rightarrow \big(T_\beta \dot{\neg} x \leftrightarrow \neg T_\beta x\big)\big]$;

(v) $\forall\beta < \alpha \forall x \forall y \big[x \dot{\wedge} y \in D^+_\beta \rightarrow \big(T_\beta(x \dot{\wedge} y) \leftrightarrow (T_\beta x \wedge T_\beta y)\big)\big]$;

(vi) $\forall\beta < \alpha \big[\dot{\forall} z.x \in D^+_\beta \rightarrow \big(T_\beta \dot{\forall} z.x \leftrightarrow \forall y T_\beta x(y)\big)\big]$;

(vii) The axiom $\mathbf{K3}_\alpha$.

Then, $\mathsf{Q}_\alpha \vdash \mathrm{TI}(\xi; U)$ for all $\xi < \varphi 1 \alpha_0(\ldots(\varphi 1 \alpha_n(\alpha | m)))$.

Since WKF_α meets all these conditions, this theorem suffices for our goal and we will give its proof in the rest of the present section.

Let $x \in tc_\beta$ denote $\mathrm{For}_\beta(x) \wedge \forall y\big(F_\beta x(y) \leftrightarrow \neg T_\beta x(y)\big)$,[11] and set:

$$\mathrm{Prog}^\beta(a) :\equiv \forall\eta\big[(\forall\xi<\eta)T_\beta a(\xi) \to T_\beta a(\eta)\big]$$

$$\mathrm{TI}^\beta(a,\gamma) :\equiv \mathrm{Prog}^\beta(a) \to \forall\xi<\gamma\ T_\beta a(\xi)$$

$$I^\beta(\gamma) :\equiv (\forall\delta<\beta)(\forall a\in tc_\delta)\ \mathrm{TI}^\delta(a,\gamma).$$

Correspondingly, given a formula ϕ, we write

$$\mathrm{Prog}(\phi) :\equiv \forall\eta\big(\forall\xi<\eta\ \phi(\xi)\to\phi(\eta)\big) \qquad \mathrm{TI}(\phi,\gamma) :\equiv \mathrm{Prog}(\phi)\to\forall\xi<\gamma\ \phi(\xi).$$

Let $\mathrm{Lim}(x)$ represent the set of limit ordinals. We can easily check these possess the following desired properties in Q_α (in fact, they are provable without using (i)-(vii)).

- $\forall\beta<\alpha\forall a\big[\forall\gamma(\mathrm{TI}^\beta(a,\gamma)\to\mathrm{TI}^\beta(a,\gamma+1))\wedge\forall\gamma{\in}\mathrm{Lim}(\forall\xi<\gamma\mathrm{TI}^\beta(a,\xi)\to\mathrm{TI}^\beta(a,\gamma))\big]$
- $\forall\beta\le\alpha\big[\forall\gamma\big(I^\beta(\gamma)\to I^\beta(\gamma+1)\big)\wedge\forall\gamma\in\mathrm{Lim}\big((\forall\xi<\gamma)I^\beta(\xi)\to I^\beta(\gamma)\big)\big]$
- $\forall\beta\le\alpha\forall\gamma\big[\forall\delta\le\beta\big(I^\beta(\gamma)\to I^\delta(\gamma)\big)\wedge\big(\beta\in\mathrm{Lim}\wedge(\forall\delta<\beta)I^\delta(\gamma)\to I^\beta(\gamma)\big)\big]$.

An $\mathcal{L}^\prec$-formula $\phi(x,\vec{z})$ is said to be *typed by* x, when ϕ is constructed from $\mathcal{L}_0$-atomics and $T_x^\prec(s)$ by logical connectives and quantifiers over variables different from x. Then, when $\phi(x,\vec{z})\in\mathcal{L}^\prec$ is typed by x, $\phi(t,\vec{z})$ is $\mathcal{L}_t^\prec$ for a closed term t with $t\in\mathrm{fd}(\prec)$ and $\forall a\in\mathrm{fd}(\prec)\big(\ulcorner\phi(\dot{a},\vec{z})\urcorner\in\mathrm{Fml}_a^\prec\big)$ (provably in $\mathsf{PA}(U)$); in particular, $\forall\beta<\alpha\big(\ulcorner\phi(\dot{\beta},\vec{z})\urcorner\in\mathrm{Fml}_\beta\big)$ when $\phi\in\mathcal{L}^{\lhd\lceil\alpha}$. We will write $\phi_a(\vec{z})$ for a typed formula $\phi(a,\vec{z})$. Notice that Prog^β and TI^β are typed by β. Now Q_α proves the following.

(a) $\forall\beta<\alpha\forall\gamma<\beta\forall a\big[\big(\mathrm{For}_\gamma(a)\to a{\in}tc_\beta\big)\wedge\big(\mathrm{St}_\gamma(a)\to D_\beta^+(a)\big)\big]$; induction on a using (i)-(iii).

[11] It stands for 'total and consistent predicate at the β-th level'; cf. [5]. It is also denoted by S_β in [43].

(b) $\forall\beta < \alpha\forall\gamma < \beta\forall\vec{z}\big[T_\beta\big(\ulcorner\phi_{\dot\gamma}(\vec{\dot z})\urcorner\big) \leftrightarrow \phi_\gamma(\vec z)\big]$, for each formula $\phi_x(\vec z)$ typed by x; by meta-induction on ϕ using (a), (i) and (iv)-(vi).

(c) $\forall\beta<\alpha\forall\gamma<\beta\forall\delta\big[\big(\mathrm{Prog}^\beta(\ulcorner\phi_{\dot\gamma}\urcorner) \leftrightarrow \mathrm{Prog}(\phi_\gamma)\big) \wedge \big(\mathrm{TI}^\beta(\ulcorner\phi_{\dot\gamma}\urcorner,\delta) \leftrightarrow \mathrm{TI}(\phi_\gamma,\delta)\big)\big]$, for each formula $\phi_x(z)$ typed by x; use (b).

(d) $\forall\beta < \alpha\forall a \in tc_\beta[\ulcorner T_{\dot\beta}\dot a(\cdot)\urcorner \in tc_\beta]$; by (vii).

(e) $\forall\beta < \alpha\forall a \in tc_\beta\big[\ulcorner\mathrm{TI}^{\dot\beta}(\dot a,\cdot)\urcorner \in tc_\beta \wedge \forall\delta\big(T_\beta\big(\ulcorner\mathrm{TI}^{\dot\beta}(\dot a,\dot\delta)\urcorner\big) \leftrightarrow \mathrm{TI}^\beta(a,\delta)\big)\big]$; all substitution instances of subformulae of $\ulcorner\mathrm{TI}^{\dot\beta}(\dot a,\cdot)\urcorner$ is in D^+_β by (d) and (i)-(iii); then the second conjunct follows from (i) and (iv)-(vii).

Lemma 5.6.3. Let $B^\beta(\gamma)$ be $\forall\delta[\forall a \in tc_\beta\mathrm{TI}^\beta(a,\delta) \to \forall a \in tc_\beta\mathrm{TI}^\beta(a,\varphi_\gamma\delta)]$. Then, we have $\mathsf{Q}_\alpha \vdash \forall\beta<\alpha\ \mathrm{Prog}\big(\lambda\gamma.B^\beta(\gamma)\big)$; shown in the standard manner (cf. [17]) using (e) above.[12]

Corollary 5.6.4. $\mathsf{Q}_\alpha \vdash \forall\beta \le \alpha\big[\beta \in \mathrm{Lim} \to \forall\xi\big(I^\beta(\xi) \to I^\beta(\varphi_\xi 0)\big)\big]$; by the last lemma and applying (c) to B^δ for $\delta<\beta$ (note that B^δ is typed by δ and thus a $tc_{\delta+1}$-predicate).

Corollary 5.6.5. $\mathsf{Q}_\alpha \vdash \forall\beta \le \alpha\big[\beta \in \mathrm{Lim} \to \mathrm{Prog}\big(\lambda\delta.I^\beta(\Gamma_\delta)\big)\big]$; use the last corollary.

The arguments so far have been parallel to [43], but we need a slight modification for the next step. Jäger et al. showed in Lemma 4 of [43] that KF_α proves

$$\forall\beta \in (\mathrm{Lim}\cap\alpha)\exists a \in tc_\beta\forall x[T_\beta a(x) \leftrightarrow I^\beta(x)], \tag{5.8}$$

which states that I^β can be represented by a tc_β-predicate a. Although they didn't give an explicit proof, such a representation a should be the code of the following kind of formula:

$$(\forall\delta<\beta)(\forall b)\big[T_\beta(\ulcorner tc_{\dot\delta}(\dot b)\urcorner) \to \mathrm{TI}^\beta(\ulcorner T_{\dot\delta}\dot b(\cdot)\urcorner, x)\big]. \tag{5.9}$$

[12] The corresponding statement for SRT_α is implicit in the proof of Lemma 5 of [43], which then corresponds to our Corollary 5.6.4 below.

For this a, since (5.9) above is an $\mathcal{L}_\beta$-sentence, KF_α can derive:

$$T_\beta a(x) \leftrightarrow (\forall \delta < \beta)(\forall b)\big[T_\beta\big(\ulcorner T_\beta(\ulcorner tc_{\dot{\delta}}(\dot{b})\urcorner)\urcorner\big) \rightarrow T_\beta\big(\ulcorner \mathrm{TI}^\beta(\ulcorner T_{\dot{\delta}}\dot{b}(\cdot)\urcorner, \dot{x})\urcorner\big)\big]; \tag{5.10}$$

thus we can focus only on $\delta < \beta$ and $b \in tc_\delta$ (note that tc_δ is typed by δ) and then obtain (5.8) in KF_α. However, (5.10) is no longer valid in WKF_α; in general, even when A is arithmetical, $T_\beta(\ulcorner A \rightarrow B\urcorner)$ isn't equivalent to $A \rightarrow T_\beta(B)$ in WKF_α. Since (5.8) is used in the proof of Main Lemma I of [43], we have to make a modification at this point.

Proposition 5.6.6. For each term $t(z, \vec{x})$, there is an $\mathcal{L}^{\prec}$-formula $\phi_z(\vec{x})$ typed by z such that

$$\mathsf{Q}_\alpha \vdash \forall \beta < \alpha \forall \gamma < \beta \forall \vec{x}\big[T_\beta(\ulcorner \phi_{\dot{\gamma}}(\dot{\vec{x}})\urcorner) \leftrightarrow I^\gamma(t(\gamma, \vec{x})\big].$$

Proof. As such ϕ, we can take $(\forall \delta < z)(\forall b)[T_z\ulcorner tc_{\dot{\delta}}(\dot{b})\urcorner \rightarrow \mathrm{TI}^z(\ulcorner T_{\dot{\delta}}\dot{b}(\cdot)\urcorner, t(z,\vec{x}))$. This ϕ is typed by z. Hence, for $\gamma < \beta < \alpha$, it follows from (b) that:

$$\forall \vec{x}\big[T_\beta\ulcorner \phi_{\dot{\gamma}}(\dot{\vec{x}})\urcorner \leftrightarrow (\forall \delta < \gamma)(\forall b)\big(T_\gamma\ulcorner tc_\delta(\dot{b})\urcorner \rightarrow \mathrm{TI}^\gamma(\ulcorner T_{\dot{\delta}}\dot{b}(\cdot)\urcorner, t(\gamma, \vec{x}))\big)\big]; \text{ and then}$$

$$\forall \vec{x}\big[T_\beta\ulcorner \phi_{\dot{\gamma}}(\dot{\vec{x}})\urcorner \leftrightarrow (\forall \delta < \gamma)(\forall b \in tc_\delta)\mathrm{TI}^\gamma(\ulcorner T_{\dot{\delta}}\dot{b}(\cdot)\urcorner, t(\gamma, \vec{x}))\big].$$

Since $\mathrm{TI}^\gamma(\ulcorner T_{\dot{\delta}}\dot{b}(\cdot)\urcorner, w) \leftrightarrow \mathrm{TI}^\delta(b, w)$ for $\delta < \gamma$ and all w (by $\mathbf{K2}_\alpha$), we obtain the claim. □

This proposition tells that I^γ can be represented at any higher level $\beta > \gamma$. In fact, the next lemma suffices for proving Main Lemma I of [43]. It follows from Proposition 5.6.6 and doesn't require for I^γ to be representable at the *same* level γ (i.e., (5.8)).

Lemma 5.6.7. $\mathsf{Q}_\alpha \vdash \forall \beta \leq \alpha \forall \gamma < \beta \forall \delta \forall \xi\big[\big(\mathrm{Lim}(\beta) \wedge I^\beta(\xi) \wedge \mathrm{Prog}(\lambda x. I^\gamma(\varphi 1 \delta x))\big) \rightarrow I^\gamma(\varphi 1 \delta \xi)\big]$.

Proof. Let $\beta \leq \alpha$ with $\mathrm{Lim}(\beta)$ and $\gamma < \beta$. Suppose $I^\beta(\xi) \wedge \mathrm{Prog}(\lambda x. I^\gamma(\varphi 1 \delta x)$. It follows from the last proposition that there is $a \in \mathrm{For}_\gamma \subset tc_{\gamma+1}$ which represents $I^\gamma(\varphi 1 \delta x)$. Then, since $\gamma + 1 < \beta$,

we obtain $\mathrm{TI}^{\gamma+1}(a,\xi)$ from $I^{\beta}(\xi)$. Finally, since $\mathrm{Prog}\big(\lambda x.I^{\gamma}(\varphi 1\delta x)\big)$ is equivalent to $\mathrm{Prog}^{\gamma+1}(a)$, we obtain $(\forall\eta<\xi)T_{\gamma+1}a(\varphi 1\delta\eta)$; thus $I^{\gamma}(\varphi 1\delta\xi)$. □

Following the notations of [43], let $\gamma\uparrow\beta$ and $\mathrm{Main}_{\alpha}(\delta)$ respectively denote

$$\exists\eta\in\mathrm{Lim}\exists\beta_0[\beta=\beta_0+\gamma\cdot\eta]\quad\text{and}\quad\forall\beta\leq\alpha\forall\gamma[\omega^{1+\delta}\uparrow\beta\wedge I^{\beta}(\gamma)\rightarrow I^{\beta}(\varphi 1\delta\gamma)],$$

Lemma 5.6.8. $\mathsf{Q}_{\alpha}\vdash\mathrm{Prog}(\lambda\delta.\mathrm{Main}_{\alpha}(\delta))$.

Proof. The proof is divided into three cases: i.e., the base case ($\delta=0$), successor case and limit case. Our proof goes parallel to Main Lemma I of [43] except that we use the last lemma instead of (5.8). We will briefly illustrate how the last lemma is used.

Assume $(\forall\eta<\delta)\mathrm{Main}_{\alpha}(\eta)$ and take $\beta\leq\alpha$ with $\omega^{1+\delta}\uparrow\beta$; then $\beta\in\mathrm{Lim}$. For a suitable fundamental sequence $\beta[n]$ ($\lim_{n\to\infty}\beta[n]=\beta$), we can show in the same way as [43] that:

$$(\forall n)\mathrm{Prog}(\lambda x.I^{\beta[n]}(\varphi 1\delta x));\tag{5.11}$$

note that we use Corollary 5.6.5 for the base case where $\delta=0$ (and IH for the other cases).

We have not yet used the last lemma until this point, and it is the next step that we need it, in which we show $(\forall n)I^{\beta[n]}(\varphi 1\delta\gamma)$ from (5.11) under the assumption $I^{\beta}(\gamma)$. This immediately follows from Lemma 5.6.7, since $\beta\in\mathrm{Lim}$ and $\beta[n]<\beta$ for all n. □

We can show the following in the standard manner:

- $\mathsf{Q}_{\alpha}\vdash\forall\beta\leq\alpha\forall\delta\forall\beta_0\big[\beta=\beta_0+\omega^{1+\delta}\rightarrow(\forall\eta<\delta)(\forall\gamma)[I^{\beta}(\gamma)\rightarrow I^{\beta}(\varphi 1\eta\gamma)]\big]$; by the last lemma and $\mathrm{TI}_{\mathcal{L}^{\lhd\uparrow\alpha}}(\alpha)$. This corresponds to Corollary 8 of [43].
- $\mathsf{Q}_{\alpha}\vdash\forall\beta\leq\alpha\forall\delta\forall\beta_0\big[\beta=\beta_0+\omega^{1+\delta}\rightarrow\mathrm{Prog}\big(\lambda\gamma.I^{\beta}(\varphi 1\delta\gamma)\big)\big]$; by Corollary 5.6.5 (for $\delta=0$) and the above (for the other cases). This corresponds to Corollary 9 of [43].

- $\mathsf{Q}_\alpha \vdash I^{\omega^{1+\alpha_0}+\cdots+\omega^{1+\alpha_n}}(\xi)$ for each $\xi < (\alpha|m)$; by Lemma 5.6.3 and the fact $\mathsf{Q}_\alpha \vdash \mathrm{TI}_{\mathcal{L}^{\lhd\uparrow\alpha}}(< \varepsilon(\alpha))$ using (a) and (c).

Using Lemma 5.6.7, these finally entail Theorem 5.6.2.

Theorem 5.6.9. $|\mathsf{WKF}_\alpha| = |\mathsf{KF}_\alpha|$ and $|\mathsf{Aut}(\mathsf{WKF})| = |\mathsf{Aut}(\mathsf{KF})|$.

5.7 Determinate Truth and Feferman Logic

The present section introduces and studies the iteration and autonomous progression of Feferman's determinate truth axiomatized by DT.

As was shown in Chapter 3, DT is identical with the system $\mathsf{FKF} + \mathrm{Cons}$. Recall that Feferman logic is a variant of weak Kleene logic in which the conditional '$\rightarrow$' is not definable in terms of '$\neg$' and '$\vee$' and the evaluation of '$\rightarrow$' is determined by the following separate truth table:

$\rightarrow$	T	U	F
T	T	U	F
U	U	U	U
F	T	T	T

The evaluations of all the other connectives and quantifiers in Feferman logic are the same as weak Kleene logic. Hence, in order to deal with the iteration of DT-truth and FKF-truth in our current framework, we assume that the conditional '$\rightarrow$' is assigned its own Gödel number independently of the other connectives (recall that we have excluded '$\rightarrow$' from the primitive logical connectives in the present chapter); thus $x \dot{\rightarrow} y \neq (\dot{\neg} x) \dot{\vee} y$. Then, the systems of iterated DT-truth and FKF-truth are defined as follows.

Definition 5.7.1. Let $\prec$ be a p.r. ordering. $D_a^{\prec}(x)$ denotes an $\mathcal{L}^{\prec}$-formula $T_a^{\prec}x \vee F_a^{\prec}x$; it expresses '$x$ is determinate at the a-th level'. The system $\mathsf{DT}_{\prec}$ over $\mathcal{L}^{\prec}$ consists of $\mathsf{PA}(U)$-axioms, full induction for $\mathcal{L}^{\prec}$, $\mathrm{TI}_{\mathcal{L}^{\prec}}(\prec)$, and:

$\mathbf{D1}_\prec$ $\forall a \in \mathrm{fd}(\prec)\big[\forall x\big(\mathrm{AtSt}_0(x) \to D_a^\prec(x)\big) \wedge \forall b \prec a \forall x\big(\mathrm{AtSt}_b^\prec(x) \to D_a^\prec(x)\big)\big]$

$\mathbf{D2}_\prec$ $\forall a, x \in \mathrm{CT}\big[a^\circ \in \mathrm{fd}(\prec) \to \big(D_{a^\circ}^\prec\big(\dot{T}_a x\big) \leftrightarrow D_a^\prec(x^\circ)\big)\big]$

$\mathbf{D3}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall x, y \in \mathrm{St}_a^\prec\big[\big(D_a^\prec(x \dot{\to} y) \leftrightarrow \big(D_a^\prec x \wedge (T_a^\prec x \to D_a^\prec y)\big)\big)$

$\wedge\big(D_a^\prec(\dot{\neg} x) \leftrightarrow D_a^\prec x\big) \wedge \big(D_a^\prec(x \dot{\wedge} y) \leftrightarrow (D_a^\prec x \wedge D_a^\prec y)\big)\big]$

$\mathbf{D4}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall z \forall x\big[\dot{\forall} z.x \in \mathrm{St}_a^\prec \to \big(D_a^\prec(\dot{\forall} z.x) \leftrightarrow \forall y D_a^\prec x(y)\big)\big]$

$\mathbf{D5}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall \vec{x} \in \mathrm{CT}[T_a^\prec(\dot{R}\vec{x}) \leftrightarrow R\vec{x}^\circ]$, for each predicate symbol R of $\mathcal{L}_0$

$\mathbf{D6}_\prec$ $\forall a, b, x \in \mathrm{CT}\big[b^\circ \preceq a^\circ \to \big((T_{a^\circ}^\prec(T_b^\prec x) \leftrightarrow T_{b^\circ}^\prec x^\circ) \wedge (T_{a^\circ}^\prec(\dot{T}_a^\prec x) \leftrightarrow T_{a^\circ}^\prec x^\circ)\big)\big]$

$\mathbf{D7}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall x, y \in D_a^\prec \cap \mathrm{St}_a^\prec\big[\big(T_a^\prec(x \dot{\to} y) \leftrightarrow (T_a^\prec x \to T_a^\prec y)\big)$

$\wedge\big(T_a^\prec(\dot{\neg} x) \leftrightarrow \neg T_a^\prec x\big) \wedge \big(T_a^\prec(x \dot{\wedge} y) \leftrightarrow (T_a^\prec x \wedge T_a^\prec y)\big)\big]$

$\mathbf{D8}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall z \forall x\big[\dot{\forall} z.x \in D_a^\prec \cap \mathrm{St}_a^\prec \to \big(T_a^\prec(\dot{\forall} z.x) \leftrightarrow \forall y T_a^\prec x(\dot{y})\big)\big]$,

Then, the system $\mathsf{Aut}(\mathsf{DT})$ is defined as $\mathsf{PA}(U)$ plus full-induction for $\mathcal{L}$, and the following two rules: for each p.r. linear ordering $\prec$ and arbitrary $\mathcal{L}$-formula ψ,

$$\frac{\mathrm{TI}(\prec; U)}{\mathbf{D1}_\prec \wedge \cdots \wedge \mathbf{D8}_\prec} \quad \text{and} \quad \frac{\mathrm{TI}(\prec; U)}{\mathrm{TI}(\prec; \psi)}.$$

Definition 5.7.2. $\mathsf{FKF}_\prec$ over $\mathcal{L}^\prec$ is defined as $\mathsf{WKF}_\prec$ plus the compositional axiom for '$\to$':

$\mathbf{F7}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall x, y \in \mathrm{St}_a^\prec$

$\big[\big(T_a^\prec(x \dot{\to} y) \leftrightarrow ((T_a^\prec x \wedge T_a^\prec y) \vee F_a^\prec x)\big) \ \wedge \ \big(F_a^\prec(x \dot{\to} y) \leftrightarrow (T_a^\prec x \wedge F_a^\prec y)\big)\big]$.

Then, $\mathsf{Aut}(\mathsf{FKF})$, $\mathsf{Aut}(\mathsf{FKF} + \mathrm{Cons})$ and $\mathsf{Aut}(\mathsf{FKF} + \mathrm{Comp})$ are defined analogously to $\mathsf{Aut}(\mathsf{KF})$, $\mathsf{Aut}(\mathsf{KF} + \mathrm{Cons})$ and $\mathsf{Aut}(\mathsf{KF} + \mathrm{Comp})$ respectively.

Proposition 5.7.3. $\mathsf{FKF}_\prec + \mathrm{Cons}_\prec$ and $\mathsf{DT}_\prec$ are identical (as theories). This is shown in the same manner as Theorem 21 of [24] with obvious modifications. Thus, $\mathsf{Aut}(\mathsf{DT})$ is also identical with $\mathsf{Aut}(\mathsf{FKF} + \mathrm{Cons})$.

Consequently, we can apply the cut-elimination technique developed in §4 to DT_α with straightforward modifications and this gives the upper bound of $|\mathsf{DT}_\alpha|$. However, we can even show that $\mathsf{DT}_\prec$ is syntactically embeddable in $\mathsf{KF}_\prec$ in general.

Lemma 5.7.4. $\mathsf{FKF}_\prec + \mathrm{Cons}_\prec$ $(= \mathsf{DT}_\prec)$ is syntactically embeddable in $\mathsf{KF}_\prec$.

Proof. Let us drop the superscript '$\prec$' from '$T^\prec$' for readability. First, we define a binary function I by the recursion theorem (or diagonalization) such that: $I(b,a) :=$

$$\begin{cases} b & \text{if } b \in \mathrm{Fml}_0 \\ \ulcorner T_c I(d,c) \wedge \neg T_c I(\dot\neg d,c) \wedge F_c I(\dot\neg d,c) \wedge \neg F_c I(d,c) \urcorner & \text{if } \exists c,d \in \mathrm{Tm}(b = \dot T_c d \wedge c^\circ \prec a) \\ \ulcorner T_c I(\dot\neg d,c) \vee \neg T_c I(d,c) \vee F_c I(d,c) \vee \neg F_c I(\dot\neg d,c) \urcorner & \text{if } \exists c,d \in \mathrm{Tm}(b = \dot\neg \dot T_c d \wedge c^\circ \prec a) \\ \ulcorner T_c I(d,c) \urcorner & \text{if } \exists c,d \in \mathrm{Tm}(b = \dot T_c d \wedge c^\circ = a) \\ \ulcorner T_c I(\dot\neg d,c) \urcorner & \text{if } \exists c,d \in \mathrm{Tm}(b = \dot\neg \dot T_c d \wedge c^\circ = a) \\ I(c,a) & \text{if } b = \dot\neg\dot\neg c \\ \big(I(c,a) \dot\wedge I(d,a) \dot\wedge \dot\neg I(\dot\neg c,a) \dot\wedge \dot\neg I(\dot\neg d,a)\big) & \text{if } b = c \dot\wedge d \\ I(\dot\neg c,a) \dot\wedge I(\dot\neg d,a) \dot\wedge \dot\neg I(c,a) \dot\wedge \dot\neg I(d,a) & \\ \quad \dot\vee \big(I(\dot\neg c,a) \dot\wedge I(d,a) \dot\wedge \dot\neg I(c,a) \dot\wedge \dot\neg I(\dot\neg d,a)\big) & \\ \quad \dot\vee \big(I(c,a) \dot\wedge I(\dot\neg d,a) \dot\wedge \dot\neg I(\dot\neg c,a) \dot\wedge \dot\neg I(d,a)\big) & \text{if } b = \dot\neg(c \dot\wedge d) \\ \big(I(c,a) \dot\wedge I(d,a) \dot\wedge \dot\neg I(\dot\neg c,a) \dot\wedge \dot\neg I(\dot\neg d,a)\big) & \\ \quad \dot\vee \big(I(\dot\neg c,a) \dot\wedge \dot\neg I(c,a)\big) & \text{if } b = c \dot\to d \\ I(c,a) \dot\wedge I(\dot\neg d,a) \dot\wedge \dot\neg I(\dot\neg c,a) \dot\wedge \dot\neg I(d,a) & \text{if } b = \dot\neg(c \dot\to d) \\ \dot\forall c.\big(I(d,a) \dot\wedge \dot\neg I(\dot\neg d,a)\big) & \text{if } b = \dot\forall c.d \\ \dot\forall c.\big[\big(I(d,a) \dot\wedge \dot\neg I(\dot\neg d,a)\big) \dot\vee \big(I(\dot\neg d,a) \dot\wedge \dot\neg I(d,a)\big)\big] & \\ \qquad \dot\wedge\ \dot\exists c.I(\dot\neg d,a) & \text{if } b = \dot\neg\dot\forall c.d; \end{cases}$$

otherwise, we set $I(b,a) = 0$. The embedding $\mathcal{T}$ of $\mathsf{DT}_\prec$ in $\mathsf{KF}_\prec$ is given by:

$$T_z x \longmapsto T_z I(x,z) \wedge \neg T_z I(\dot\neg x,z) \wedge F_z I(\dot\neg x,z) \wedge \neg F_z I(x,z).$$

$\mathcal{T}$ is of essentially the same form as the syntactical embedding of DT in KF given in [24] and the proof is its straightforward generalization except the case for $\mathbf{K2}_\prec$. Thus, we only illustrate the

case for $\mathbf{K2}_\prec$: i.e., $\mathsf{KF}_\prec \vdash (\mathbf{K2}_\prec)^{\mathcal{T}}$. For $b, x \in \mathrm{CT}$ with $b^\circ \prec a$,

$$\begin{aligned}
&\big(F_a(\dot{T}_b x)\big)^{\mathcal{T}} \leftrightarrow T_a I(\dot{\neg}\dot{T}_b x, a) \wedge \neg T_a I(\dot{T}_b x, a) \wedge F_a I(\dot{T}_b x, a) \wedge \neg F_a I(\dot{\neg}\dot{T}_b x, a)\\
&\leftrightarrow \big[T_a{\ulcorner}T_b I(\dot{\neg}x, b){\urcorner} \vee T_a{\ulcorner}\neg T_b I(x, b){\urcorner} \vee T_a{\ulcorner}F_b I(x, b){\urcorner} \vee T_a{\ulcorner}\neg F_b I(\dot{\neg}x, b){\urcorner}\big]\\
&\quad \wedge \big[\neg T_a{\ulcorner}T_b I(x, b){\urcorner} \vee \neg T_a{\ulcorner}\neg T_b I(\dot{\neg}x, b){\urcorner} \vee \neg T_a{\ulcorner}F_b I(\dot{\neg}x, b){\urcorner} \vee \neg T_a{\ulcorner}\neg F_b I(x, b){\urcorner}\big]\\
&\quad \wedge \big[F_a{\ulcorner}T_b I(x, b){\urcorner} \vee F_a{\ulcorner}\neg T_b I(\dot{\neg}x, b){\urcorner} \vee F_a{\ulcorner}F_b I(\dot{\neg}x, b){\urcorner} \vee F_a{\ulcorner}\neg F_b I(x, b){\urcorner}\big]\\
&\quad \wedge \big[\neg F_a({\ulcorner}T_b I(\dot{\neg}x, b){\urcorner} \vee \neg F_a{\ulcorner}\neg T_b I(x, b){\urcorner} \vee \neg F_a({\ulcorner}F_b I(x, b){\urcorner} \vee \neg F_a{\ulcorner}\neg F_b I(\dot{\neg}x, b){\urcorner}\big]\\
&\leftrightarrow T_{b^\circ} I(\dot{\neg}x^\circ, b^\circ) \vee \neg T_{b^\circ} I(x^\circ, b^\circ) \vee F_{b^\circ} I(x^\circ, b^\circ) \vee \neg F_{b^\circ} I(\dot{\neg}x^\circ, b^\circ) \equiv \big(\neg T_{b^\circ}(x^\circ)\big)^{\mathcal{T}};
\end{aligned}$$

note that each of the four conjuncts of the third formula is equivalent to the fourth. We can dually show $(T_a{\ulcorner}T_b x{\urcorner})^{\mathcal{T}} \leftrightarrow (T_b x^\circ)^{\mathcal{T}}$ in the parallel manner. □

We can easily show that $\mathsf{WKF}_\prec$+$\mathrm{Comp}_\prec$ and $\mathsf{KF}_\prec$+$\mathrm{Comp}_\prec$ are identical; cf. Lemma 51 of [24]. It is also observed that, for all $a \in \mathrm{fd}(\prec)$ and $x, y \in \mathrm{St}_a$, $\mathsf{FKF}_\prec$+$\mathrm{Comp}_\prec$ derives $T_a(x \dot{\rightarrow} y) \leftrightarrow T_a(\dot{\neg}x \dot{\vee} y)$ and $F_a(x \dot{\rightarrow} y) \leftrightarrow F_a(\dot{\neg}x \dot{\vee} y)$. Namely, in the presence of $\mathrm{Comp}_\prec$, the inner logics of WKF and FKF collapse into strong Kleene logic. Now, by the recursion theorem, we can take a function h such that $h({\ulcorner}T_a t{\urcorner}) = {\ulcorner}T_a h(t){\urcorner}$, $h({\ulcorner}\phi \to \psi{\urcorner}) = \dot{\neg}h({\ulcorner}\phi{\urcorner}) \dot{\vee} h({\ulcorner}\psi{\urcorner})$ and commutes with the other connectives. Then, the embedding of $\mathsf{FKF} + \mathrm{Comp}_\prec$ in $\mathsf{KF} + \mathrm{Comp}_\prec$ is given by $T_z x \mapsto T_z h(x)$. The main theorem of this section is thus obtained:

Theorem 5.7.5. (1) $|\mathsf{DT}_\alpha| = |\mathsf{FKF}_\alpha| = |\mathsf{WKF}_\alpha + \mathrm{Cons}_\alpha \text{ (or } \mathrm{Comp}_\alpha)| = |\mathsf{KF}_\alpha|$.

(2) $|\mathsf{Aut}(\mathsf{DT})| = |\mathsf{Aut}(\mathsf{FKF})| = |\mathsf{Aut}(\mathsf{WKF}+\mathrm{Cons})| = |\mathsf{Aut}(\mathsf{WKF}+\mathrm{Comp})| = \varphi 200$.

5.8 Cantini's VF-Truth

In the present section, we consider the iteration of this VF-truth and its autonomous progression.

5.8.1 The Systems $\mathsf{VF}_\prec$ and $\mathsf{Aut}(\mathsf{VF})$

Given a system S, Ax_S represents the set of the axioms of S and Prv_S is a canonical provability predicate of S. Let $\mathsf{PL}_{\mathcal{L}^\prec}$ be pure classical predicate logic for the language $\mathcal{L}^\prec$ (thus with equality); then $\mathrm{Ax}_{\mathsf{PL}_{\mathcal{L}^\prec}}$ represents the set of logical axioms (for $\mathcal{L}^\prec$). Our following formulation of $\mathsf{VF}_\prec$ is based on the simpler version of VF in [7] (or [45]) rather than the original one in [6].

Definition 5.8.1. Given a p.r. ordering $\prec$, the system $\mathsf{VF}_\prec$ over $\mathcal{L}^\prec$ comprises $\mathsf{PA}(U)$, full-induction for $\mathcal{L}^\prec$, $\mathrm{TI}_{\mathcal{L}^\prec}(\prec)$ and the following truth axioms:

$\mathbf{V1}_\prec$ $\forall\vec{x}\forall a \in \mathrm{fd}(\prec)[T_a^\prec(\ulcorner\phi(\dot{\vec{x}})\urcorner) \to \phi(\vec{x})]$, for each $\mathcal{L}^\prec$-formula $\phi(\vec{x})$.

$\mathbf{V2}_\prec$ $\forall a{\in}\mathrm{fd}(\prec)\forall\vec{x}{\in}\mathrm{CT}[(T_a^\prec(\dot{R}\vec{x}){\leftrightarrow}R\vec{x}^\circ){\wedge}(F_a^\prec(\dot{R}\vec{x}){\leftrightarrow}{\neg}R\vec{x}^\circ)]$, for each $R{\in}\mathcal{L}_0$

$\mathbf{V3}_\prec$ $\forall a{\in}\mathrm{fd}(\prec)\forall b,x{\in}\mathrm{CT}[b^\circ{\prec}a \to (T_a^\prec(\dot{T}_b^\prec x) \leftrightarrow T_{b^\circ}^\prec x^\circ){\wedge}(F_a^\prec(\dot{T}_b^\prec x) \leftrightarrow {\neg}T_{b^\circ}^\prec x^\circ)]$

$\mathbf{V4}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall x \in \mathrm{St}^\prec[\mathrm{Ax}_{\mathrm{PL}_{\mathcal{L}^\prec}}(x) \to T_a^\prec(x)]$

$\mathbf{V5}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall z\forall x[\mathrm{St}^\prec(\dot{\forall}z.x) \to (\forall y T_a^\prec x(y) \to T_a^\prec\dot{\forall}z.x)]$

$\mathbf{V6}_\prec$ $\forall a \in \mathrm{fd}(\prec)\forall x,y \in \mathrm{St}^\prec[T_a^\prec(x\dot{\to}y) \to (T_a^\prec x \to T_a^\prec y)]$

The system $\mathsf{Aut}(\mathsf{VF})$ is defined as $\mathsf{PA}(U)$ plus full-induction for $\mathcal{L}$, and the following rules:

$$\frac{\mathrm{TI}(\prec;U)}{\phi}, \qquad \frac{\mathrm{TI}(\prec;U)}{\mathbf{V2}_\prec \wedge \cdots \wedge \mathbf{V6}_\prec} \quad \text{and} \quad \frac{\mathrm{TI}(\prec;U)}{\mathrm{TI}(\prec;\psi)},$$

for each p.r. linear ordering $\prec$, instance ϕ of $\mathbf{V1}_\prec$ and $\mathcal{L}$-formula ψ.

Proposition 5.8.2. The following are all provable in $\mathsf{VF}_\prec$.

(i) $\forall a \in \mathrm{fd}(\prec)\forall x \in \mathrm{St}^\prec\big[\mathrm{Prv}_{\mathrm{PL}_{\mathcal{L}^\prec}}(x) \to T_a^\prec(x)\big]$.

(ii) $\forall a \in \mathrm{fd}(\prec)\forall x,y \in \mathrm{St}^\prec$

$\big[\big(T_a^\prec(x\dot{\wedge}y) \leftrightarrow (T_a^\prec x \wedge T_a^\prec y)\big) \wedge \big((T_a^\prec x \vee T_a^\prec y) \to T_a^\prec(x\dot{\vee}y)\big)\big]$.

(iii) $\forall a \in \mathrm{fd}(\prec)\forall x\forall z$

$$\big[\dot\forall z.x \in \mathrm{St}^{\prec} \to \big((\forall y T_a^{\prec} x(y) \leftrightarrow T_a^{\prec}\dot\forall z.x) \wedge (\exists y T_a^{\prec} x(y) \to T_a^{\prec}\dot\exists z.x)\big)\big].$$

(iv) $\forall a \in \mathrm{fd}(\prec)\forall x \in \mathrm{St}^{\prec}[T_a^{\prec} x \leftrightarrow T_a^{\prec}(\dot\neg\dot\neg x)]$.

(v) $\forall x \in \mathrm{For}^{\prec}\forall z \in \mathrm{CT}[T_a^{\prec}\mathrm{sb}(x;z) \leftrightarrow T_a^{\prec}\mathrm{sb}(x;\mathrm{nm}\circ\mathrm{val}(z))]$.

(vi) For each $\mathcal{L}^{\prec}$-formula $\phi_x(\vec z)$ typed by x (see §6 for its definition),

$$\forall a{\in}\mathrm{fd}(\prec)\forall b,\vec z{\in}\mathrm{CT}[b^{\circ}{\prec}a \to (T_a^{\prec}(\ulcorner\phi_b(\vec z)\urcorner) \leftrightarrow \phi_{b^{\circ}}\vec{z^{\circ}}) \wedge (F_a^{\prec}(\ulcorner\phi_b(\vec z)\urcorner) \leftrightarrow \neg\phi_{b^{\circ}}\vec{z^{\circ}})].$$

We notice that (i) says 'every logically valid sentence is true at any level' and (v) expresses 'the truth of a formula $A(t)$ depends only on the value of the closed term t (at any level)'.

Proof. We can show (i)-(v) in a parallel manner to Proposition 2.1 of [6]. For (vi), we show the both conjuncts simultaneously by meta-induction on ϕ; one direction immediately follows from $\mathbf{V1}_{\prec}$ using (v); the converse is shown by $\mathbf{V2}_{\prec}$ and $\mathbf{V3}_{\prec}$ for the base step and by (ii) and (iii) for the induction steps. □

In what follows, we first embed the system $\mathsf{ID}_{\prec}$ of iterated inductive definitions along $\prec$ in $\mathsf{VF}_{\prec}$ and then embed $\mathsf{VF}_{\prec}$ in the system Π^1_1-$\mathsf{CA}_{\prec}$ + (Bi) of iterated Π^1_1-comprehensions along $\prec$ plus the Bar induction (Bi) (with the proviso that $\prec$ has no end point); thus we have $\mathsf{ID}_{\prec} \subset \mathsf{VF}_{\prec} \subset \Pi^1_1$-$\mathsf{CA}_{\prec}$ + (Bi). Fix a natural notation system up to a sufficiently large ordinal suitable for ordinal analyses of impredicative systems. Then, together with already known results on ID_{α} and Π^1_1-CA_{α} (and also KPI_{α}) [3, 12, 65], these syntactical embeddings entail that VF_{α} is proof-theoretically equivalent to ID_{α} and Π^1_1-CA_{α} + (Bi) for a limit α; in the same ordinal notation as in [3] and [13], we also have

$$|\mathsf{VF}_{\alpha}| = \Theta\varepsilon_{\Omega_{\alpha}+1}0 \ \text{ (for a limit } \alpha) \quad \text{and} \quad |\mathsf{Aut}(\mathsf{VF})| = \Theta\Omega_{\Omega_1}0.$$

As in the preceding sections, we drop the superscript $\prec$ in what follows for readability whenever there is no worry of confusion.

5.8.2 Lower Bound

Kahle [45] gave a syntactical embedding of ID_1 in VF. In the present subsection, we generalize it to an embedding of $\mathsf{ID}_\prec$ in $\mathsf{VF}_\prec$.

Given a p.r. ordering $\prec$, the language $\mathcal{L}_{\mathsf{ID}}$ of $\mathsf{ID}_\prec$ is $\mathcal{L}_0$ plus new unary predicates $J^{\mathcal{A}}$ associated to each inductive operator form $\mathcal{A}$. We write $J^{\mathcal{A}}_a(b)$ and $J^{\mathcal{A}}_{\prec a}(b)$ for $J^{\mathcal{A}}(\langle b, a\rangle)$ and $(b)_1 \prec a \wedge J^{\mathcal{A}}(b)$ respectively. The theory $\mathsf{ID}_\prec$ consists of $\mathsf{PA}(U)$, full induction for $\mathcal{L}_{\mathsf{ID}}$, $\mathrm{TI}_{\mathcal{L}_{\mathsf{ID}}}(\prec)$ and the axiom schemata of inductive definitions:

$(\mathbf{ID}^1_\prec)$ $\forall a \in \mathrm{fd}(\prec)\forall x[\mathcal{A}(J^{\mathcal{A}}_a, J^{\mathcal{A}}_{\prec a}, x, a) \to J^{\mathcal{A}}_a(x)]$.

$(\mathbf{ID}^2_\prec)$ $\forall a \in \mathrm{fd}(\prec)\big[\forall x\big(\mathcal{A}(\psi, J^{\mathcal{A}}_{\prec a}, x, a) \to \psi(x)\big) \to \forall x\big(x \in J^{\mathcal{A}}_a \to \psi(x)\big)\big]$,

for arbitrary $\mathcal{L}_{\mathsf{ID}}$-formula ψ.

Let $\mathcal{A}(R, Q, x, y)$ be an inductive operator form, and let $\psi(x)$, $\theta(x)$, $\chi(x)$ be $\mathcal{L}^\prec$-formulae possibly with other parameters $\vec{v}$. By straightforward meta-induction on $\mathcal{A}$, we can show:

$$\mathsf{PL}_{\mathcal{L}^\prec} \vdash \forall\vec{v}\forall y\forall x\big[\big(\mathcal{A}(\psi, \theta, x, y) \wedge \forall z(\psi(z) \to \chi(z))\big) \to \mathcal{A}(\chi, \theta, x, y)\big]. \tag{5.12}$$

Lemma 5.8.3. Let $\theta_x(z, \vec{v})$ be a $\mathcal{L}^\prec$-formula typed by x with a distinguished variable z and parameters $\vec{v}$ and let $\theta_{\prec x}(z, \vec{v})$ denote $(z)_1 \prec x \wedge \theta_{(z)_1}(z, \vec{v})$. Then, $\mathsf{VF}_\prec$ derives:

$$\forall\vec{v}\forall a{\in}\mathrm{fd}(\prec)\forall z\big[\big(T_a{}^\ulcorner\theta_{\prec\dot{a}}(\dot{z}, \dot{\vec{v}})^\urcorner \leftrightarrow \theta_{\prec a}(z, \vec{v})\big) \wedge \big(F_a{}^\ulcorner\theta_{\prec\dot{a}}(\dot{z}, \dot{\vec{v}})^\urcorner \leftrightarrow \neg\theta_{\prec a}(z, \vec{v})\big)\big]$$

Proof. Fix $a \in \mathrm{fd}(\prec)$ and $\vec{v}$. For both conjuncts, the one direction immediately follows from $\mathbf{V1}_\prec$ and the converse follows from $\mathbf{V2}_\prec$ and Proposition 5.8.2 (ii) and (vi). □

Lemma 5.8.4. Let $\mathcal{A}(R,Q,x,y)$ be an inductive operator form and let $\psi(x,z)$ and $\theta(x,z)$ be $\mathcal{L}^{\prec}$-formulae possibly with parameters $\vec{v}$. Suppose $\mathsf{VF}_{\prec}$ derives:

$$\forall\vec{v}\forall a{\in}\mathrm{fd}(\prec)\forall z\big[\big(T_a\ulcorner\theta(\dot{a},\dot{z},\dot{\vec{v}})\urcorner\leftrightarrow\theta(a,z,\vec{v})\big)\wedge\big(F_a\ulcorner\theta(\dot{a},\dot{z},\dot{\vec{v}})\urcorner\leftrightarrow\neg\theta(a,z,\vec{v})\big)\big]. \tag{5.13}$$

We write $\psi_{a,\vec{v}}(z)$ and $\theta_{a,\vec{v}}(z)$ for $\psi(a,z,\vec{v})$ and $\theta(a,z,\vec{v})$ respectively. Then,

$$\mathsf{VF}_{\prec}\vdash\forall\vec{v}\forall a{\in}\mathrm{fd}(\prec)\forall z\big[\mathcal{A}\big(\lambda u.T_a\ulcorner\psi_{\dot{a},\dot{\vec{v}}}(\dot{u})\urcorner,\ \theta_{a,\vec{v}},z,a\big)\rightarrow T_a\ulcorner\mathcal{A}(\psi_{\dot{a},\vec{v}},\ \theta_{\dot{a}\dot{\vec{v}}},\dot{z},\dot{a})\urcorner\big].$$

Proof. Meta-induction on $\mathcal{A}$; use $\mathbf{V2}_{\prec}$, (5.13) and Proposition 5.8.2. □

We introduce two notations which generalize $\mathrm{Clos}_{\mathcal{A}}(\psi)$ and ψ' in [45]:

$$\begin{aligned}\mathrm{Clos}_{\mathcal{A}}[\psi;\theta]\ &:\equiv\ \forall x\big[\mathcal{A}(\psi,\theta,x,y)\rightarrow\psi(x)\big]\\ \psi'_\theta(x)\ &:\equiv\ \mathrm{Clos}_{\mathcal{A}}[\psi;\theta]\rightarrow\psi(x)\end{aligned}$$

Lemma 5.8.5. Let $\mathcal{A}(R,Q,x,y)$ be an inductive operator form and $\psi(x)$ and $\theta(x)$ be arbitrary $\mathcal{L}^{\prec}$-formulae possibly with parameters y and $\vec{v}$. Then, $\mathsf{PL}_{\mathcal{L}^{\prec}}\vdash(\forall y)(\forall\vec{v})\mathrm{Clos}_{\mathcal{A}}[\psi'_\theta;\theta]$.

Proof. A straightforward generalization of Lemma 33 of [45]. We have to show that

$$\forall y\forall\vec{v}\forall x\big[\mathcal{A}\big(\psi'_\theta,\theta,x,y\big)\rightarrow\big(\mathrm{Clos}_{\mathcal{A}}[\psi;\theta]\rightarrow\psi(x)\big)\big]$$

Fix the parameters. Take any x and suppose $\mathcal{A}(\psi'_\theta,\theta,x,y)$ and $\mathrm{Clos}_{\mathcal{A}}[\psi;\theta]$. By $\mathrm{Clos}_{\mathcal{A}}[\psi;\theta]$, we have $\forall z[\psi'_\theta(z)\rightarrow\psi(z)]$. Then, it follows from (5.12) that $\mathcal{A}(\psi'_\theta,\theta,x,a)\rightarrow\mathcal{A}(\psi,\theta,x,y)$; and thus $\mathcal{A}(\psi,\theta,x,y)$. Again by $\mathrm{Clos}_{\mathcal{A}}[\psi;\theta]$, we finally obtain $\psi(x)$. □

In what follows, we suppress all irrelevant parameters for simplicity.

Lemma 5.8.6. Let $\mathcal{A}(R,Q,x,y)$ be an inductive operator form and $\theta_z(x)$ be the same as Lemma 5.8.4. Then, $\mathsf{VF}_\prec \vdash \forall a \in \mathrm{fd}(\prec)\ \mathrm{Clos}_{\mathcal{A}}[T_a{\ulcorner}\psi'_{\theta_{\dot a}}{\urcorner};\theta_a]$, for any $\mathcal{L}^\prec$-formula ψ.

Proof. It follows from the last lemma and Proposition 5.8.2 (i) that

$$\mathsf{VF}_\prec \vdash \forall a \in \mathrm{fd}(\prec)\ T_a{\ulcorner}\mathrm{Clos}_{\mathcal{A}}[\psi'_{\theta_{\dot a}};\theta_{\dot a}]{\urcorner}; \text{i.e.,}$$

$$\mathsf{VF}_\prec \vdash \forall a \in \mathrm{fd}(\prec)\ T_a{\ulcorner}\forall x(\mathcal{A}(\psi'_{\theta_{\dot a}},\theta_{\dot a},x,\dot a) \to \psi'_{\theta_{\dot a}}(x)){\urcorner}; \text{by Prop 5.8.2(iii) and } \mathbf{V6}_\prec,$$

$$\mathsf{VF}_\prec \vdash \forall a \in \mathrm{fd}(\prec)\ \forall x[T_a{\ulcorner}\mathcal{A}(\psi'_{\theta_{\dot a}},\theta_{\dot a},\dot x,\dot a){\urcorner} \to T_a{\ulcorner}\psi'_{\theta_{\dot a}}(\dot x){\urcorner}].$$

We also have $\forall a{\in}\mathrm{fd}(\prec)\forall x[\mathcal{A}(T_a{\ulcorner}\psi'_{\theta_{\dot a}}{\urcorner},\theta_a,x,a) \to T_a{\ulcorner}\mathcal{A}(\psi'_{\theta_{\dot a}},\theta_{\dot a},\dot x,\dot a){\urcorner}]$ by Lemma 5.8.4. Combining these, we obtain $\forall a{\in}\mathrm{fd}(\prec)\forall x[\mathcal{A}(T_a{\ulcorner}\psi'_{\theta_{\dot a}}{\urcorner},\theta_a,x,a) \to T_a{\ulcorner}\psi'_{\theta_{\dot a}}(\dot x){\urcorner}]$. □

Given an inductive operator form $\mathcal{A}(R,Q,x,y)$, let $\Phi^{\mathcal{A}}(x,y,w)$ denote:

$$\forall z \in \mathrm{For}^\prec\Big(\mathrm{Clos}_{\mathcal{A}}\big[T_y z(\cdot);\lambda u.\big((u)_1 \prec y \wedge T_{(u)_1} w((u)_0,(u)_1)\big)\big] \to T_y z(x)\Big).$$

By diagonalization (on w), there is a closed term t such that $t = {\ulcorner}\Phi^{\mathcal{A}}(x,y,t){\urcorner}$ and thus $\Phi^{\mathcal{A}}(x,y,t) \leftrightarrow \Phi^{\mathcal{A}}\big(x,y,{\ulcorner}\Phi^{\mathcal{A}}(x,y,t){\urcorner}\big)$. Set $S^{\mathcal{A}}(x,y) :\equiv \Phi^{\mathcal{A}}(x,y,t)$ and $P^{\mathcal{A}}(x,y) :\equiv T_y\big({\ulcorner}S^{\mathcal{A}}(\dot x,\dot y){\urcorner}\big)$. Let $\Psi^{\mathcal{A}}(u,y)$ be $T_y t((u)_0,(u)_1)$. Then $\Psi^{\mathcal{A}}(u,y)$ is typed by y and $\Psi^{\mathcal{A}}_{\prec y}(u) \equiv (u)_1 \prec y \wedge T_{(u)_1} t((u)_0,(u)_1)$ in the notation of Lemma 5.8.3. By definition, we have

$$S^{\mathcal{A}}(x,y) \ \equiv\ \forall z \in \mathrm{For}^\prec\big(\mathrm{Clos}_{\mathcal{A}}[T_y z(\cdot);\lambda u.\Psi^{\mathcal{A}}_{\prec y}(u)] \to T_y z(x)\big) \tag{5.14}$$

$$P^{\mathcal{A}}(x,y) \ \equiv\ T_y\Big({\ulcorner}\forall z \in \mathrm{For}^\prec\big(\mathrm{Clos}_{\mathcal{A}}[T_{\dot y} z(\cdot);\lambda u.\Psi^{\mathcal{A}}_{\prec \dot y}(u)] \to T_{\dot y} z(\dot x)\big){\urcorner}\Big). \tag{5.15}$$

We define an embedding * of $\mathsf{ID}_\prec$ in $\mathsf{VF}_\prec$ by $J^{\mathcal{A}}(x) \mapsto P^{\mathcal{A}}((x)_0, (x)_1)$. Then,

$$(J^{\mathcal{A}}_{\prec y}x)^\star \equiv (x)_1 \prec y \wedge P^{\mathcal{A}}((x)_0, (x)_1) \tag{5.16}$$

$$\equiv (x)_1 \prec y \wedge T_{(x)_1}\ulcorner S^{\mathcal{A}}((\dot{x})_0, (\dot{x})_1)\urcorner \tag{5.17}$$

$$\leftrightarrow (x)_1 \prec y \wedge T_{(x)_1} t((x)_0, (x)_1)\urcorner \equiv \Psi^{\mathcal{A}}_{\prec y}(x). \tag{5.18}$$

Let us write $P^{\mathcal{A}}_{\prec y}x$ for $(x)_1 \prec y \wedge P^{\mathcal{A}}((x)_0, (x)_1)$, i.e., $(J^{\mathcal{A}}_{\prec y}x)^\star$. By (5.16), we have

$$\mathcal{A}^\star(J^{\mathcal{A}}_y, J^{\mathcal{A}}_{\prec y}, x, y) \equiv \mathcal{A}(P^{\mathcal{A}}_y, P^{\mathcal{A}}_{\prec y}, x, y) \leftrightarrow \mathcal{A}(P^{\mathcal{A}}_y, \Psi^{\mathcal{A}}_{\prec y}, x, y) \tag{5.19}$$

The following two lemmata will establish that the translation * is indeed an embedding.

Lemma 5.8.7. $\mathsf{VF}_\prec \vdash \forall a \in \mathrm{fd}(\prec)\forall x\big[\mathcal{A}(P^{\mathcal{A}}_a, P^{\mathcal{A}}_{\prec a}, x, a) \to P^{\mathcal{A}}_a(x)\big]$, for each inductive operator form $\mathcal{A}(R, Q, x, y)$.

Proof. Fix arbitrary $a \in \mathrm{fd}(\prec)$. By logic, it follows from (5.14) that

$$\mathsf{PL}_{\mathcal{L}^\prec} \vdash \forall x\forall z \in \mathrm{For}^\prec\big[S^{\mathcal{A}}_a(x) \to \big(\mathrm{Clos}_{\mathcal{A}}[T_a z(\cdot); \Psi^{\mathcal{A}}_{\prec a}] \to T_a z(x)\big)\big]; \text{ and thus}$$

$$\mathsf{PL}_{\mathcal{L}^\prec} \vdash \forall x\forall z \in \mathrm{For}^\prec\big[\mathrm{Clos}_{\mathcal{A}}[T_a z(\cdot); \Psi^{\mathcal{A}}_{\prec a}] \to \big(S^{\mathcal{A}}_a(x) \to T_a z(x)\big)\big].$$

Hence, it follows from (5.12) that, in $\mathsf{PL}_{\mathcal{L}^\prec}$,

$$\forall x\forall z \in \mathrm{For}^\prec\big[\mathrm{Clos}_{\mathcal{A}}[T_a z(\cdot); \Psi^{\mathcal{A}}_{\prec a}] \to \big(\mathcal{A}(S^{\mathcal{A}}_a, \Psi^{\mathcal{A}}_{\prec a}, x, a) \to \mathcal{A}(T_a z(\cdot), \Psi^{\mathcal{A}}_{\prec a}, x, a)\big)\big].$$

Since z is not free in $\mathcal{A}(S^{\mathcal{A}}_a, \Psi^{\mathcal{A}}_{\prec a}, x, a)$, we equivalently have: in $\mathsf{PL}_{\mathcal{L}^\prec}$,

$$\forall x\big[\mathcal{A}(S^{\mathcal{A}}_a, \Psi^{\mathcal{A}}_{\prec a}, x, a) \to \forall z \in \mathrm{For}^\prec\big(\mathrm{Clos}_{\mathcal{A}}[T_a z(\cdot); \Psi^{\mathcal{A}}_{\prec a}] \to \mathcal{A}(T_a z(\cdot), \Psi^{\mathcal{A}}_{\prec a}, x, a)\big)\big].$$

Recall that $\mathrm{Clos}_{\mathcal{A}}[T_a z(\cdot); \Psi^{\mathcal{A}}_{\prec a}]$ denotes $\forall x\big[\mathcal{A}(T_a z(\cdot), \Psi^{\mathcal{A}}_{\prec a}, x, a) \to T_a z(x)\big]$; thus,

$$\mathsf{PL}_{\mathcal{L}^{\prec}} \vdash \forall x\big[\mathcal{A}(S^{\mathcal{A}}_a, \Psi^{\mathcal{A}}_{\prec a}, x, a) \to \forall z \in \mathrm{For}^{\prec}\big(\mathrm{Clos}_{\mathcal{A}}[T_a z(\cdot); \Psi^{\mathcal{A}}_{\prec a}] \to T_a z(x)\big)\big].$$

Since all these have been proved in $\mathsf{PL}_{\mathcal{L}^{\prec}}$, $\mathsf{VF}_{\prec}$ proves that, for all x,

$$T_a{}^{\ulcorner}\mathcal{A}(S^{\mathcal{A}}_{\dot{a}}, \Psi^{\mathcal{A}}_{\prec \dot{a}}, \dot{x}, \dot{a})^{\urcorner} \to T_a{}^{\ulcorner}\forall z \in \mathrm{For}^{\prec}\big(\mathrm{Clos}_{\mathcal{A}}[T_{\dot{a}} z(\cdot); \Psi^{\mathcal{A}}_{\prec \dot{a}}] \to T_{\dot{a}} z(\dot{x})\big)^{\urcorner}. \tag{5.20}$$

On the other hand, since $\Psi^{\mathcal{A}}_a$ is typed, it follows from Lemmata 5.8.3 and 5.8.4,

$$\mathcal{A}(P^{\mathcal{A}}_a, \Psi^{\mathcal{A}}_{\prec a}, x, a) \equiv \mathcal{A}(T_a{}^{\ulcorner}S^{\mathcal{A}\urcorner}_{\dot{a}}, \Psi^{\mathcal{A}}_{\prec a}, x, a) \to T_a{}^{\ulcorner}\mathcal{A}(S^{\mathcal{A}}_{\dot{a}}, \Psi^{\mathcal{A}}_{\prec \dot{a}}, \dot{x}, \dot{a})^{\urcorner}. \tag{5.21}$$

Thus, our claim finally follows from (5.19), (5.21), (5.20) and (5.14). □

Lemma 5.8.8. For each inductive operator $\mathcal{A}(R, Q, x, y)$ and an arbitrary $\mathcal{L}^{\prec}$-formula $\psi(x)$,

$$\mathsf{VF}_{\prec} \vdash \forall a \in \mathrm{fd}(\prec)\big[\forall x\big(\mathcal{A}(\psi, P^{\mathcal{A}}_{\prec a}, x, a) \to \psi(x)\big) \to \forall y\big(P^{\mathcal{A}}_a(y) \to \psi(y)\big)\big].$$

Proof. Fix $a \in \mathrm{fd}(\prec)$ and take any y. Suppose $P^{\mathcal{A}}_a(y)$. By $\mathbf{V1}_{\prec}$, we obtain $S^{\mathcal{A}}_a(y)$ and thus $\forall z \in \mathrm{For}^{\prec}\big(\mathrm{Clos}_{\mathcal{A}}[T_a z(\cdot); P^{\mathcal{A}}_{\prec a}] \to T_a z(y)\big)$ by (5.16). Let ϕ_a be $\psi'_{P^{\mathcal{A}}_{\prec a}}$; then we have

$$\mathrm{Clos}_{\mathcal{A}}\big[T_a\big({}^{\ulcorner}\phi_{\dot{a}}(\cdot)^{\urcorner}\big); P^{\mathcal{A}}_{\prec a}\big] \to T_a\big({}^{\ulcorner}\phi_{\dot{a}}(\dot{y})^{\urcorner}\big). \tag{5.22}$$

Since $P^{\mathcal{A}}_a$ is typed by a, we already have $\mathrm{Clos}_{\mathcal{A}}\big[T_a\big({}^{\ulcorner}\phi_{\dot{a}}(\cdot)^{\urcorner}\big); P^{\mathcal{A}}_{\prec a}\big]$ by Lemmata 5.8.3 and 5.8.6. Then, (5.22) entails $T_a\big({}^{\ulcorner}\phi_{\dot{a}}(\dot{y})^{\urcorner}\big)$ and thus $\phi_a(y)$ again by $\mathbf{V1}_{\prec}$. Finally, $\psi(y)$ follows under the condition $\forall x\big(\mathcal{A}(\psi, P^{\phi}_{\prec a}, x, a) \to \psi(x)\big)$ by the definition of ϕ_a $(\equiv \psi'_{P^{\mathcal{A}}_{\prec a}})$. □

Theorem 5.8.9. We can regard $\mathsf{ID}_{\prec} \subset \mathsf{VF}_{\prec}$. In particular, the system $\overline{\mathsf{ID}}$ of autonomously iterated

inductive definitions (see [3]) is embeddable in $\mathsf{Aut}(\mathsf{VF})$.

5.8.3 Upper Bound

In the present subsection, we embed $\mathsf{VF}_{\prec}$ in Π^1_1-$\mathsf{CA}_{\prec}$ + (Bi). Let $\mathcal{L}^2$ be the second-order language of arithmetic obtained from $\mathcal{L}_0$; we assume that the classes Π^1_n (Σ^1_n) of $\mathcal{L}^2$-formulae are suitably defined so that $\Pi^1_0 = \Sigma^1_0$ (i.e., the class of 'arithmetical' formulae) includes $\mathcal{L}_0$. Given a p.r. ordering $\prec$, the axiom schema Π^1_1-CA$_{\prec}$ is defined by:

$$\forall z \in \mathrm{fd}(\prec)\exists X\forall y \prec z\forall x[x \in X^y \leftrightarrow \phi(X^{\prec y}, x, y)],$$

for each Π^1_1-formula $\phi(Z, x, y)$ only with the displayed parameters, where $x \in X^y$ and $x \in X^{\prec y}$ respectively denote $\langle x, y\rangle \in X$ and $(x)_1 \prec y \wedge x \in X$. Then, the system Π^1_1-$\mathsf{CA}_{\prec}$ over $\mathcal{L}^2$ is a subsystem of second-order arithmetic defined as Π^1_1-CA plus $\mathrm{Wf}(\prec) := \forall X[\mathrm{Prog}(\prec; X) \to \forall x \in \mathrm{fd}(\prec)(x \in X)\big]$ ('$\prec$ is well-founded') and Π^1_1-CA$_{\prec}$. We notice that this definition of Π^1_1-$\mathsf{CA}_{\prec}$ is different from the one in [64].[13] The Bar induction (Bi) is defined in the standard manner; see, [64]. Feferman [13] showed that (Bi) is equivalent, modulo ACA, to the so-called quantifier schema:

(QS) $\forall X\phi(X) \to \phi[\psi/X]$, for any arithmetical ϕ and arbitrary ψ.

A Π^1_1-formula $\mathcal{A}(X, Y, x, y)$ possibly with other parameters is called a *monotone* Π^1_1 operator (with respect to X), when it holds that

$$\forall X\forall X'\big[X \subset X' \to \big(\mathcal{A}(X, Y, x, y) \to \mathcal{A}(X', Y, x, y)\big)\big].$$

In particular, when $\mathcal{A}(X, Y, x, y)$ is X-positive arithmetical, it is monotone Π^1_1.

[13] It is observed from Theorem 3.2.4.2 and the proof of Lemma 3.3.5.2 of [64] that $\mathsf{KPI}_{\alpha} \vdash \Pi^1_1$-CA$_{\alpha}$. It is also known that $\mathsf{KPI} \vdash \Pi^1_1$-CA + (Bi); their proofs are found in [64]. Hence, Π^1_1-CA_{α} + (Bi) in our definition can be embedded in KPI_{α}. Then, the proof-theoretic equivalence of ID_{α} and KPI_{α} (Theorem 3.4.6.18 of [64]) finally yields our claim.

Given a monotone Π^1_1 operator $\mathcal{A}(X, Y, x, y)$, define

$$\begin{aligned} \mathrm{Clos}_{\mathcal{A}}(X, Y, y) &:\equiv (\forall x)[\mathcal{A}(X, Y, x, y) \to x \in X]. \\ H_{\mathcal{A}}(Y, x, y) &:\equiv (\forall Z)[\mathrm{Clos}_{\mathcal{A}}(Z, Y, y) \to x \in Z] \end{aligned}$$

Informally, $H_{\mathcal{A}}(Y, x, y)$ expresses $x \in \bigcap\{Z \mid \mathrm{Clos}_{\mathcal{A}}(Z, Y, y)\}$. $H_{\mathcal{A}}$ doesn't appear to be Π^1_1 at a first glance. However, we can show that $\mathrm{Clos}_{\mathcal{A}}$ is Σ^1_1 in Π^1_1-CA using Σ^1_1-AC ($\dashv \Pi^1_1$-CA); therefore $H_{\mathcal{A}}$ is Π^1_1 in Π^1_1-CA. Let S be the set defined by $H_{\mathcal{A}}$. Then, we can show in the standard manner that S is the least fixed-point of $\mathcal{A}$; see [13] or [64].

When $\mathcal{A}(X, Y, x, y)$ contains no other parameters, we can apply Π^1_1-$\mathrm{CA}_{\prec}$ to $H_{\mathcal{A}}$ and obtain:

$$\Pi^1_1\text{-}\mathsf{CA}_{\prec} \vdash \forall z \in \mathrm{fd}(\prec)(\exists X)(\forall y \prec z)(\forall x)[x \in X^y \leftrightarrow H_{\mathcal{A}}(X^{\prec y}, x, y)].$$

Given $z \in \mathrm{fd}(\prec)$, let S be a set such that $(\forall x)(\forall y \prec z)[x \in S^y \leftrightarrow H_{\mathcal{A}}(S^{\prec y}, x, y)]$. Then, for each $y \prec z$, we have

$$(\forall X)[\mathrm{Clos}_{\mathcal{A}}(X, S^{\prec y}, y) \to S^y \subset X] \quad \text{and} \quad \forall x[\mathcal{A}(S^y, S^{\prec y}, x, y) \to x \in S^y].$$

These imply $\forall x[\mathcal{A}(S^y, S^{\prec y}, x, y) \leftrightarrow x \in S^y]$ ('S^y is a fixed-point of $\mathcal{A}(\ \cdot, S^{\prec y}, x, y)$').

Let us fix a p.r. ordering $\prec$. We will construct an embedding of $\mathsf{VF}_{\prec}$ in Π^1_1-$\mathsf{CA}_{\prec}$ + (Bi). As was shown by Cantini [6], VF can be modelled by a Kripkean fixed-point with van Fraassen's supervaluation schema. Indeed, $\mathsf{VF}_{\prec}$ can be similarly modelled by iterating such structures. In what follows, we will formalize this argument within Π^1_1-$\mathsf{CA}_{\prec}$ + (Bi).

Given $V \subset \mathbb{N}$, let $(\mathbb{N}, V)$ be a $\mathcal{L}^{\prec}$-structure in which the truth predicate $T^{\prec}$ is interpreted by $(V)_0 \times (V)_1$ where $(V)_i = \{(x)_i \mid x \in V\}$ for $i = 0, 1$; we simply write $V \models \phi$ for $(\mathbb{N}, V) \models \phi$ for each $\mathcal{L}^{\prec}$-sentence ϕ. It is known that the set $\{\sigma \in \mathcal{L}^{\prec} \mid (\mathbb{N}, V) \models \sigma\}$ is Π^1_1 (indeed Δ^1_1) in V; see [6].

This fact can be formalized in Π^1_1-CA. We define an Z-positive arithmetical formula $\mathcal{B}(Z, X, Y, x, a)$ as follows:

$$\bigwedge_{R\in\mathcal{L}_0} \exists \vec{y} \in \mathrm{CT}\big[\big(x = \dot{R}\vec{y} \wedge R(\vec{y}^{\circ})\big) \vee \big(x = \dot{\neg}\dot{R}\vec{y} \wedge \neg R(\vec{y}^{\circ})\big)\big]$$

$$\vee\, \exists b, y \in \mathrm{CT}\big[b^{\circ} \prec a \wedge [(x = \dot{T}_b y \wedge y^{\circ} \in Y^{b^{\circ}}) \vee (x = \dot{\neg}\dot{T}_b y \wedge y^{\circ} \notin Y^{b^{\circ}})]\big]$$

$$\vee\, \exists b, y \in \mathrm{CT}\big[b^{\circ} \not\prec a \wedge [(x = \dot{T}_b y \wedge y^{\circ} \in X^{b^{\circ}}) \vee (x = \dot{\neg}\dot{T}_b y \wedge y^{\circ} \notin X^{b^{\circ}})]\big]$$

$$\vee\, \exists y, z \in \mathrm{St}^{\prec}\big[(x = y \dot{\wedge} z \wedge y \in Z \wedge z \in Z) \vee \big(x = \dot{\neg}(y \dot{\wedge} z) \wedge (\dot{\neg} y \in Z \vee \dot{\neg} z \in Z)\big)\big]$$

$$\vee\, \exists y \exists z \big[\big(x \in \mathrm{St}^{\prec} \wedge x = \dot{\forall} y.z \wedge \forall v(z(v) \in Z)\big)$$

$$\vee \big(x \in \mathrm{St}^{\prec} \wedge x = \dot{\neg}\dot{\forall} y.z \wedge \exists v(\dot{\neg} z(v) \in Z)\big)\big]$$

$$\vee\, \exists y \in \mathrm{St}^{\prec}(x = \dot{\neg}\dot{\neg} y \wedge y \in Z).$$

Let $\mathrm{Tr}(X, Y, x, a)$ be the Π^1_1-formula defining the least fixed-point of $\mathcal{B}$ (i.e., $H_{\mathcal{B}}(X, Y, x, a)$). We sometimes write $x \in \mathrm{Tr}(X, Y, a)$ for $\mathrm{Tr}(X, Y, x, a)$. As we shall see below, $\mathrm{Tr}(X, Y, x, a)$ represents the set $\{\ulcorner \phi \urcorner \mid (X^{\not\prec a} \cup Y^{\prec a}) \models \phi\}$ as is expected, where $X^{\not\prec a}$ denotes the set $\{\langle x, y\rangle \in X \mid y \not\prec a\}$ $(\supset X^a)$.

The next is shown by using the fixed-point property of $\mathrm{Tr}(X, Y, a)$.

Proposition 5.8.10. The following hold in Π^1_1-CA: for all X, Y and $a \in \mathrm{fd}(\prec)$,

(1) $\forall \vec{x} \in \mathrm{CT}\big[\big(\dot{R}\vec{x} \in \mathrm{Tr}(X,Y,a) \leftrightarrow R\vec{x}^{\circ}\big) \wedge \big(\dot{\neg}\dot{R}\vec{x} \in \mathrm{Tr}(X,Y,a) \leftrightarrow \neg R\vec{x}^{\circ}\big)\big]$, for $R \in \mathcal{L}_0$

(2) $\forall x, b \in \mathrm{CT}\big[b^{\circ} \prec a \rightarrow \big[\big(\dot{T}_b x \in \mathrm{Tr}(X, Y, a) \leftrightarrow x^{\circ} \in Y^{b^{\circ}}\big)$

$\wedge\big(\dot{\neg}\dot{T}_b x \in \mathrm{Tr}(X, Y, a) \leftrightarrow x^{\circ} \notin Y^{b^{\circ}}\big)\big]\big]$

(3) $\forall x, b \in \mathrm{CT}\big[b^{\circ} \not\prec a \rightarrow \big[\big(\dot{T}_b x \in \mathrm{Tr}(X, Y, a) \leftrightarrow x^{\circ} \in X^{b^{\circ}}\big)$

$\wedge\big(\dot{\neg}\dot{T}_b x \in \mathrm{Tr}(X, Y, a) \leftrightarrow x^{\circ} \notin X^{b^{\circ}}\big)\big]\big]$

(4) $\forall x, y \in \mathrm{St}^{\prec}\big[[x \dot{\wedge} y \in \mathrm{Tr}(X,Y,a) \leftrightarrow (x \in \mathrm{Tr}(X,Y,a) \wedge y \in \mathrm{Tr}(X,Y,a))]$

$\wedge\big[\dot{\neg}(x\dot{\wedge}y) \in \mathrm{Tr}(X,Y,a) \leftrightarrow (\dot{\neg}x \in \mathrm{Tr}(X,Y,a) \vee \dot{\neg}y \in \mathrm{Tr}(X,Y,a))\big]\big]$

(5) $\forall x \forall z\big[\dot{\forall} z.x \in \mathrm{St}^{\prec} \rightarrow \big[\dot{\forall} z.x \in \mathrm{Tr}(X,Y,a) \leftrightarrow \forall y\big(x(y) \in \mathrm{Tr}(X,Y,a)\big)\big]$

$\wedge\big[\dot{\neg}\dot{\forall} z.x \in \mathrm{Tr}(X,Y,a) \leftrightarrow \exists y\big(\dot{\neg}x(y) \in \mathrm{Tr}(X,Y,a)\big)\big]\big]$

(6) $\forall x \in \mathrm{St}^{\prec}\big(\dot{\neg}\dot{\neg}x \in \mathrm{Tr}(X,Y,a) \leftrightarrow x \in \mathrm{Tr}(X,Y,a)\big)$

(7) $\forall x \in \mathrm{St}^{\prec}\big(\dot{\neg}x \in \mathrm{Tr}(X,Y,a) \leftrightarrow x \notin \mathrm{Tr}(X,Y,a)\big)$; by induction on x using (1)-(6).

For an $\mathcal{L}^{\prec}$-formula ϕ and a set M, ϕ^M will denote the $\mathcal{L}^2$-formula obtained from ϕ by replacing each occurrence of $T_z^{\prec}(x)$ by $x \in M^z$. Then the next proposition follows from the last one by meta-induction on ϕ.

Proposition 5.8.11. For each $\mathcal{L}^{\prec}$-formula $\phi(\vec{x})$, we have

$$\Pi^1_1\text{-}\mathsf{CA} \vdash \forall X \forall Y \forall a \forall \vec{x}[\ulcorner \phi(\vec{\dot{x}}) \urcorner \in \mathrm{Tr}(X,Y,a) \leftrightarrow \phi^{(X^{\not\prec a} \cup Y^{\prec a})}(\vec{x})].$$

We can formalize the soundness of $\mathsf{PL}_{\mathcal{L}^{\prec}}$ (i.e., soundness of classical logic):

Proposition 5.8.12. $\Pi^1_1\text{-}\mathsf{CA} \vdash \forall X \forall Y \forall a \forall x \in \mathrm{St}^{\prec}\big(\mathrm{Ax}_{\mathsf{PL}_{\mathcal{L}^{\prec}}}(x) \rightarrow x \in \mathrm{Tr}(X,Y,a)\big)$.

Define an operator $\mathrm{FV}(X,Y,x,a)$ (on x with the parameters X, Y and a) by

$$\mathrm{FV}(X,Y,x,a) \ :\equiv\ \forall Z[X \subset Z^a \rightarrow x \in \mathrm{Tr}(Z,Y,a)].$$

$\mathrm{FV}(X,Y,x,a)$ informally expresses the set

$$\{\ulcorner \phi \urcorner \mid \phi \in \mathcal{L}^{\prec} \text{ and } (Z^{\not\prec a} \cup Y^{\prec a}) \models \phi \text{ for all } Z \text{ s.t. } Z^a \supset X\ \}.$$

Since Tr is Π^1_1 and $\mathrm{FV}(X,Y,x,a)$ is X-positive, $\mathrm{FV}(X,Y,x,a)$ is a monotone Π^1_1 operator (w.r.t. X). Hence, we can define the least fixed-point $H_{\mathrm{FV}}(Y,x,a)$ of FV:

$$\begin{aligned}\mathrm{Clos}_{\mathrm{FV}}(X,Y,a) &:\equiv (\forall x)[\mathrm{FV}(X,Y,x,a) \to x \in X]\\ H_{\mathrm{FV}}(Y,x,a) &:\equiv (\forall Z)[\mathrm{Clos}_{\mathrm{FV}}(Z,Y,a) \to x \in Z].\end{aligned}$$

Recall that $\mathrm{Clos}_{\mathrm{FV}}(X,Y,a)$ is Σ^1_1 and $H_{\mathrm{FV}}(Y,x,a)$ is Π^1_1. In what follows, we assume that $\prec$ has no end point, i.e., $\forall a \in \mathrm{fd}(\prec)\exists b \in \mathrm{fd}(\prec)(a \prec b)$; therefore, if $\prec$ is well-ordered then $otyp(\prec)$ is a limit ordinal. Then, we have

$$\Pi^1_1\text{-}\mathsf{CA}_{\prec} \vdash \forall a \in \mathrm{fd}(\prec)\exists M \forall b \preceq a \forall x \big[x \in M^b \leftrightarrow H_{\mathrm{FV}}(M^{\prec b}, x, b)\big]. \tag{5.23}$$

Now, we define a translation * from $\mathcal{L}^{\prec}$ to $\mathcal{L}^2$ as follows:

$$T_z x \;\longmapsto\; \exists M \big[\forall w \preceq z \forall y \big(y \in M^w \leftrightarrow H_{\mathrm{FV}}(M^{\prec w}, y, w)\big) \wedge x \in M^z\big].$$

Our goal is to show that * is indeed an embedding of $\mathsf{VF}_{\prec}$ in $\Pi^1_1\text{-}\mathsf{CA}_{\prec} + (\mathrm{Bi})$.

Let $\mathrm{VF}_{\prec}(z,M)$ denote a formula $\forall w \preceq z \forall x[x \in M^w \leftrightarrow H_{\mathrm{FV}}(M^{\prec w}, x, w)]$; thus (5.23) is rewritten as $\Pi^1_1\text{-}\mathsf{CA}_{\prec} \vdash \forall a \in \mathrm{fd}(\prec)\exists M \mathrm{VF}_{\prec}(a,M)$. For simplicity, we make the following abuses of notation: for each $\mathcal{L}^{\prec}$-formula ϕ,

$$\phi^{\prec a}(x) :\equiv (x)_1 \prec a \wedge \phi((x)_0) \qquad \& \qquad \phi^{\not\prec a}(x) :\equiv (x)_1 \not\prec a \wedge \phi((x)_0).$$

Proposition 5.8.13. One can straightforwardly verify the following; cf. [13, 64].

$$\Pi^1_1\text{-}\mathsf{CA} + \mathrm{Wf}(\prec) \vdash \forall a \in \mathrm{fd}(\prec)\forall M\forall N$$
$$\big[\mathrm{VF}_\prec(a, M) \wedge \mathrm{VF}_\prec(a, N) \to (\forall b \preceq a)\forall x[x \in M^b \leftrightarrow x \in N^b].$$

Lemma 5.8.14. $\Pi^1_1\text{-}\mathsf{CA}_\prec + (\mathrm{Bi}) \vdash (\mathbf{V1}_\prec)^*$.

Proof. Suppose $(T_a{\ulcorner}\phi(\vec{x}){\urcorner})^*$ for $a \in \mathrm{fd}(\prec)$ and $\phi \in \mathcal{L}^\prec$. Take N with $\mathrm{VF}_\prec(a, N)$. Then, by Proposition 5.8.13, we have ${\ulcorner}\phi(\vec{x}){\urcorner} \in N^a$. Since N^a is a fixed-point of $\mathrm{FV}(\ \cdot, N^{\prec a}, a)$,

$$(\forall Z)[N^a \subset Z^a \to {\ulcorner}\phi(\vec{x}){\urcorner} \in \mathrm{Tr}(Z, N^{\prec a}, a)].$$

By Proposition 5.8.11, we obtain

$$(\forall Z)[N^a \subset Z^a \to \phi^{(Z^{\not\prec a} \cup N^{\prec a})}(\vec{x})].$$

Since the part '$N^a \subset Z^a \to \phi^{(Z^{\not\prec a} \cup N^{\prec a})}(\vec{x})$' is arithmetical, by (QS), we obtain:

$$N^a \subset ((\lambda y.Ty)^*)^a \to \phi^{\left(((\lambda x.Tx)^*)^{\not\prec a} \cup N^{\prec a}\right)}(\vec{x});$$

here, we are treating $(\lambda y.Ty)^*$ as if it were a set, but it is in fact a mere abbreviation. Then, since $N^a = \big((\lambda y.Ty)^*\big)^a$ by Proposition 5.8.13, we obtain

$$\phi^{\left(((\lambda x.Tx)^*)^{\not\prec a} \cup N^{\prec a}\right)}(\vec{x}).$$

Since $((\lambda y.Ty)^*)^{\not\prec a} \cup N^{\prec a} = (\lambda y.Ty)^*$ by Proposition 5.8.13, we obtain $\phi^*(\vec{x})$. □

Proposition 5.8.15. $\Pi^1_1\text{-}\mathsf{CA}_\prec \vdash (\mathbf{V3}_\prec)^* \wedge (\mathbf{V2}_\prec)^*$.

Proof. We only show $(\mathbf{V3}_\prec)^*$. $(\mathbf{V2}_\prec)^*$ is shown similarly. Given $a \in \mathrm{fd}(\prec)$, let N be such that $\mathrm{VF}(a, N)$ and take $b, x \in \mathrm{CT}$ with $b^\circ \prec a$. Suppose $(T_a \dot{T}_b x)^*$. Then, by Proposition 5.8.13, we have $\dot{T}_b x \in N^a$. Since N^a is a fixed-point of $\mathrm{FV}(\ \cdot, N^{\prec a}, a)$, we have

$$(\forall Z)[N^a \subset Z^a \to \dot{T}_b x \in \mathrm{Tr}(Z, N^{\prec a}, a)].$$

In particular, we have $\dot{T}_b x \in \mathrm{Tr}(N^a, N^{\prec a}, a)$. By Proposition 5.8.10-(2), we have $x^\circ \in N^{b^\circ}$ and thus $(T_{b^\circ} x^\circ)^*$. For the converse, suppose $(T_{b^\circ} x^\circ)^*$. Then $x^\circ \in N^{b^\circ}$ by Proposition 5.8.13 and $\forall Z[\dot{T}_b x \in \mathrm{Tr}(Z, N^{\prec a}, a)]$ by the definition of Tr; in particular, we have $\dot{T}_b x \in \mathrm{FV}(N^a, N^{\prec a}, a)$. Since N^a is a fixed-point, we have $\dot{T}_b x \in N^a$ and thus $(T_a \dot{T}_b x)^*$. The other conjunct is shown similarly. □

Proposition 5.8.16. $\Pi^1_1\text{-}\mathsf{CA}_\prec \vdash (\mathbf{V4}_\prec)^*$.

Proof. Take $a \in \mathrm{fd}(\prec)$ and suppose $\mathrm{Ax}_{\mathrm{PL}_{\mathcal{L}^\prec}}(x)$. Let N be such that $\mathrm{VF}(a, N)$. It suffices to show that $x \in N^a$; this follows from Proposition 5.8.12. □

Proposition 5.8.17. $\Pi^1_1\text{-}\mathsf{CA}_\prec \vdash (\mathbf{V5}_\prec)^* \wedge (\mathbf{V6}_\prec)^*$.

Proof. We only show $(\mathbf{V6}_\prec)^*$; $(\mathbf{V5}_\prec)^*$ is similarly shown. Let $a \in \mathrm{fd}(\prec)$ and $x \dot{\to} y \in \mathrm{St}^\prec$. Suppose $(T_a x \dot{\to} y)^*$ and $(T_a x)^*$. Let N be such that $\mathrm{VF}(a, N)$. By Proposition 5.8.13, we have $x, x \dot{\to} y \in N^a$. Then, since N^a is a fixed-point of FV, $(\forall Z)[N^a \subset Z^a \to x, x \dot{\to} y \in \mathrm{Tr}(Z, N^{\prec a}, a)]$. By Proposition 5.8.10, for all Z with $Z^a \supset N^a$,

$$x \in \mathrm{Tr}(Z, N^{\prec a}, a) \to y \in \mathrm{Tr}(Z, N^{\prec a}, a) \quad \text{and} \quad x \in \mathrm{Tr}(Z, N^{\prec a}, a);$$

therefor $y \in \mathrm{Tr}(Z, N^{\prec a}, a)$. Hence we have $\mathrm{FV}(N^a, N^{\prec a}, y, a)$. Again since N^a is fixed-point, we obtain $y \in N^a$. This implies $(T_a y)^*$. □

Theorem 5.8.18. Let $\prec$ be a p.r. ordering with no end point. Then, $\mathsf{VF}_{\prec}$ is syntactically embeddable in Π^1_1-$\mathsf{CA}_{\prec} + (\mathrm{Bi})$.

5.9 Summary and Concluding Remarks of Chapter 5

Let us summarize the results obtained in the present chapter.

$$\begin{aligned}|\mathsf{KF}_\alpha| &= |\mathsf{KF}_\alpha + \mathrm{Cons}_\alpha \text{ (or } \mathrm{Comp}_\alpha)| = |\mathsf{WKF}_\alpha| = |\mathsf{WKF}_\alpha + \mathrm{Cons}_\alpha \text{ (or } \mathrm{Comp}_\alpha)| \\ &= |\mathsf{PUTB}_\alpha| = |\mathsf{PUTB}_\alpha + \mathrm{Cons}_\alpha \text{ (or } \mathrm{Comp}_\alpha)| \\ &= |\mathsf{FKF}_\alpha| = |\mathsf{DT}_\alpha \ (= \mathsf{FKF}_\alpha + \mathrm{Cons}_\alpha)| = |\mathsf{FKF}_\alpha + \mathrm{Comp}_\alpha| = |\widehat{\mathsf{ID}}_\alpha|;\end{aligned}$$

in fact, they are all proof-theoretically equivalent. We also have

$$\begin{aligned}|\mathsf{Aut}(\mathsf{WKF})| &= |\mathsf{Aut}(\mathsf{WKF} + \mathrm{Cons})| = |\mathsf{Aut}(\mathsf{WKF} + \mathrm{Comp})| \\ &= |\mathsf{Aut}(\mathsf{PUTB})| = |\mathsf{Aut}(\mathsf{PUTB} + \mathrm{Cons})| = |\mathsf{Aut}(\mathsf{PUTB} + \mathrm{Comp})| \\ &= |\mathsf{Aut}(\mathsf{FKF})| = |\mathsf{Aut}(\mathsf{DT})| = |\mathsf{Aut}(\mathsf{FKF} + \mathrm{Comp})| \\ &= |\mathsf{Aut}(\mathsf{KF} + \mathrm{Cons})| = |\mathsf{Aut}(\mathsf{KF} + \mathrm{Comp})| = |\mathsf{Aut}(\mathsf{KF})| = \varphi 200;\end{aligned}$$

they are all proof-theoretically equivalent as well. Also, for a limit α,

$$|\mathsf{VF}_\alpha| = |\mathsf{ID}_\alpha| = |\Pi^1_1\text{-}\mathsf{CA}_\alpha + \mathrm{Bi}| = \Theta\varepsilon_{\Omega_\alpha+1}0 \quad \text{and} \quad |\mathsf{Aut}(\mathsf{VF})| = \Theta\Omega_{\Omega_1}0.$$

There are at least two directions for further studies from what we have obtained. First, we may further investigate the iterations and autonomous progressions of other self-applicable truths. Second, as far as I know, the system $\mathsf{Aut}(\mathsf{VF})$ is the strongest system ever among the so far presented truth systems which have natural and philosophical motivations behind them. However, its proof-

theoretic strength is far below that of, e.g., Π^1_2-CA. Seeking for a well-motivated and stronger system of self-applicable truth might be of some significance from the foundational point of view.

Chapter 6

Further Prospects

I conclude my thesis by making a very brief comment on the results and discussions so far developed.

I have given proof-theoretic analyses to a variety of axiomatic systems of truth and compared them by some different means. I dare not deduce from them a hasty conclusion about which system is the best or most adequate system of truth. The research I have so far conducted is not of the kind by which one can directly and immediately deduce such a conclusion. Rather, as I stated at the end of Chapter 1, I would like to suggest that we now need to conduct tedious but down-to-earth 'field work' of examining various systems of truth from various points of view. As long as we have decided to renounce our very naïve conception of truth, we have lost general guiding principles for constructing systems of truth and it seems to me unlikely that we can obtain the answer to 'what is *the* system of truth' merely by the traditional kind of philosophical speculations. Nonetheless, I still hope and would like to claim that my results have provided philosophers with more information on the respective notions of truth and on how they are related and compared, by which one might be able to select some systems as good and sieve out others as bad for their purpose.

The axiomatic study of truth is relatively a young area of research, which has been attracting more and more interest from philosophers and logicians in the last two decades. I expect that the

axiomatic study of truth will open a new rich and fruitful scientific field of study and I hope that my research in the present thesis will contribute to it. Instead of drawing a definite conclusion, I raise some further prospects and perspectives for future studies of axiomatic truth in the rest of this short chapter.

Since Feferman [17] and Friedman and Sheard [23], many subsequent researchers have been following the same arithmetical setting that they adopted and the study of axiomatic truths has so far been concentrated mainly on those over arithmetic; the present thesis of mine also follows this tradition and setting.[1] Accordingly, axiomatic systems of truth have been a target of traditional proof theory and ordinal analysis of arithmetic. However, we can of course consider another setting and adopt a different kind of base system over which truth is formulated. The most natural alternative is probably set theory. It might even be argued that set theory is a more suitable setting for investigating the notion of mathematical truth, since set theory is often considered to be the foundation of mathematics. However, not much attention has so far been paid to this direction of study and, as far as I know, no research paper has been published on this topic. There are some plausible reasons for having put set theory aside from the axiomatic study of truth. First, from a foundational point of view, while set theory is often considered to be a sufficiently rich framework within which we can develop contemporary mathematics, it is sometimes suspected to be too strong and doubted to be necessary as the overall framework for mathematics. For these reasons, some people do not accept strong set theories such as ZF. However, there are still many who accept ZF or its equivalents and take it as the foundation of mathematics, and thus it may well be of a good significance to study axiomatic truths over set theory and to investigate what would be brought about by them. Second, from a practical and technical point of view, the study of axiomatic truth systems over ZF is subsumed in that of the realm below inaccessibles (but above ZF), to which

[1] In fact, Feferman [17] suggests possible formulations of truth systems over set theory as well, although his study there is principally devoted to those over arithmetic.

not much attention has been paid and for which not enough proof-theoretic techniques have been developed. As far as the I see, any so far presented (natural) system of truth for arithmetic, when suitably reformulated for ZF, cannot reach the strength of the postulate of the existence of inaccessible cardinals; for, they have a 'standard' model over $\mathbb{N}$ (with some exceptions like Friedman and Sheard's FS [23, 31]) and thus, when they are formulated over ZF, we can construct their 'standard' models over the set V_κ for an inaccessible κ in completely parallel manners. However, some new perspectives and research interests about this realm below inaccessibles have recently been proposed and versatile proof-theoretic techniques have consequently been developed by Jäger and his colleagues [40, 41, 42, 44] which turn out to enable us to analyze axiomatic truth systems for set theory.[2] With these backgrounds, it may well be a good time to start the axiomatic study of truth for set theory.

Second, of course, we may well push forward the axiomatic study of truth over arithmetic under various directions. I suggest two directions among them: I propose to search for and study metapredicative and impredicative systems of truth. As for the former, the notion of metapredicativity was recently introduced by Jäger, Strahm and their colleagues. Roughly speaking, metapredicative systems are those which go beyond the predicative realm but can be analysed by predicative methods without resorting to the typical techniques in impredicative proof theory such as the technique of collapsing functions. Some metapredicative systems of truth have been already presented by in Chapter 5. However, these systems are of iterated truth and we haven't yet known any natural metapredicative system of non-iterated single truth. I also expect that the study of metapredicative truth will contribute to better understanding of metapredicativity and its connection with impred-

[2] Jäger's initial motivation seems to have come from his analysis of Feferman's operational set theory OST. He showed in [41] that a certain extension of OST is equivalent to the subsystem $\mathsf{NBG}_{<E_0}$ of MK (cf. §3.3). The study of class theory along this line is kept vigorously developing by Jäger and other logicians. Besides these recent approaches to class theory by Feferman, Jäger and their colleagues, Sato's recent works on his system SS [73] bring another perspective on class theory. Sato proposed the system SS as a versatile framework covering a broad range of mathematical logic from bounded arithmetic to set theory even beyond ZF. He showed that SS plus the existence of an inaccessible cardinal (Inacc) is equivalent to ZF (by private communication). In SS, its axiom schemata are restricted to fixed bounded complexities; then if we strengthen them by admitting more complex formulae, the resulting systems together with Inacc become stronger than ZF and correspond to some subsystems of MK.

icativity. As for impredicative truth, we already have Cantini's system VF [6] of truth. We have also seen that the impredicative system $\mathsf{Aut}(\mathsf{VF})$ is fairly strong and goes beyond Δ^1_2-CA; to repeat, $\mathsf{Aut}(\mathsf{VF})$ is the strongest system of truth so far to my knowledge, but it still falls far short of, say, Π^1_2-CA. According to my view explained in Ch.5-§1, finding a strong but natural system of truth contributes to the justification of richer mathematics. I thus propose to pursue stronger but natural impredicative systems of truth.

Bibliography

[1] Jon Barwise and Lawrence Moss. *Vicious Circle*. CSLI Publications, Stanford, 1996.

[2] Lev Beklemishev. Proof-theoretic analysis by iterated reflection. *Archive for Mathematical Logic*, 42:515–552, 2003.

[3] Wilfried Buchholz and Wolfram Pohlers. Provable wellorderings of formal theories for transfinitely iterated inductive definitions. *The Journal of Symbolic Logic*, 43:118–125, 1978.

[4] Andrea Cantini. A note on predicatively reducible theory of iterated elementary induction. *Bollettino Unione Mathematica Italiana*, 4-B 6:413–430, 1985.

[5] Andrea Cantini. Notes on formal theories of truth. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 35:97–130, 1989.

[6] Andrea Cantini. A theory of formal truth arithmetically equivalent to ID1. *The Journal of Symbolic Logic*, 55:244–259, 1990.

[7] Andrea Cantini. *Logical Frameworks for Truth and Abstraction*. Elsevier, Amsterdam, 1996.

[8] Cezary Cieśliński. Deflationism, conservativeness and maximality. *Journal of Philosophical Logic*, 36:695–705, 2007.

[9] Solomon Feferman. Reflecting on incompleteness. Handwritten Notes, March 1987, 39 pages.

[10] Solomon Feferman. Systems of predicative analysis. *The Journal of Symbolic Logic*, 29:1–30, 1964.

[11] Solomon Feferman. Systems of predicative analysis II: Representations of ordinals. *The Journal of Symbolic Logic*, 33:193–220, 1968.

[12] Solomon Feferman. Formal theories for transfinite iteration of generalized inductive definitions and some subsystems of analysis. In Akiko Kino, John Myhill, and Richard Vesley, editors, *Intuitionism and Proof Theory*, pages 303–326, Amsterdam, 1970. North-Holland.

[13] Solomon Feferman. Formal theories for transfinite iteration of generalized inductive definitions and some subsystems of analysis. In Akiko Kino, John Myhill, and Richard Vesley, editors, *Intuitionism and Proof Theory*, pages 303–326, Amsterdam, 1970. North-Holland.

[14] Solomon Feferman. Iterated inductive fixed-point theories: Application to hancock's conjecture. In George Metakides, editor, *Patras Logic Symposion*, pages 171–196. North-Holland, Amsterdam, 1982.

[15] Solomon Feferman. Toward useful type-free theories. *Journal of Symbolic Logic*, 49:75–111, 1984.

[16] Solomon Feferman. Hilbert's program relativized: Proof-theoretical and foundational reductions. *The Journal of Symbolic Logic*, 53:364–384, 1988.

[17] Solomon Feferman. Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56:1–49, 1991.

[18] Solomon Feferman. Does reductive proof theory have a viable rationale? *Erkenntnis*, 53:63–96, 2000.

[19] Solomon Feferman. Axioms for determinateness and truth. *The Review of Symbolic Logic*, 1:204–217, 2008.

[20] Solomon Feferman. Operational set theory and small large cardinals. *Annals of Pure and Applied Logic*, 207:971–979, 2009.

[21] Hartly Field. Deflating the conservativeness argument. *The Journal of Philosophy*, 96:533–540, 1999.

[22] Hartly Field. A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, 32:139–177, 2003.

[23] Harvey Friedman and Michael Sheard. An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33:1–21, 1987.

[24] Kentaro Fujimoto. Relative truth definability of axiomatic theories of truth. *The Bulletin of Symbolic Logic*, 16:305–344, 2010.

[25] Kurt Gödel. Russell's mathematical logic. In Paul Benacerraf and Hilary Putnam, editors, *Philosophy of Mathematics*, pages 447–469. Cambridge University Press, Cambridge, 1983.

[26] Anil Gupta and Nuel Belnap. *The Revision Theory of Truth*. MIT Press, Cambridge, Massachusetts, 1993.

[27] Petr Hájek. *Metamathematics of Fuzzy Logic*. Kluwer, Dordrecht, 1998.

[28] Petr Hájek, Jeffrey Paris, and John Shepherdson. The liar paradox and fuzzy logic. *The Journal of Symbolic Logic*, 65:339–346, 2000.

[29] Petr Hájek and Pavel Pudlak. *Metamathematics of First-Order Arithmetic*. Springer, Berlin, 1993.

[30] Volker Halbach. Axiomatic theories of truth. In Edward Zalta, editor, *Stanford Encyclopedia of Philosophy*. URL: http://plato.stanford.edu/entries/truth-axiomatic.

[31] Volker Halbach. A system of complete and consistent truth. *Notre Dame Journal of Formal Logic*, 35:311–327, 1994.

[32] Volker Halbach. *Axiomatische Wahrheitstheorien*. Akademie Verlag, Berlin, 1996.

[33] Volker Halbach. Conservative theories of classical truth. *Studia Logica*, 62:353–370, 1999.

[34] Volker Halbach. Disquotationalism and infinite conjunctions. *Mind*, 108:1–22, 1999.

[35] Volker Halbach. How innocent is deflationism? *Synthese*, 126:167–194, 2001.

[36] Volker Halbach. Reducing compositional to disquotational truth. *The Review of Symbolic Logic*, 2:786–798, 2009.

[37] Volker Halbach. *Axiomatic Theories of Truth*. Cambridge University Press, 2010.

[38] Volker Halbach and Leon Horsten. Axiomatizing Kripke's theory of truth. *The Journal of Symbolic Logic*, 71:677–712, 2006.

[39] Paul Horwhich. The minimalist conception of truth. In Simon Blackburn and Keith Simmons, editors, *Truth*, pages 239–263. Oxford University Press, Oxford, 1999.

[40] Gerhard Jäger. On Feferman's operational set theory OST. *Annals of Pure and Applied Logic*, 150:19–39, 2007.

[41] Gerhard Jäger. Full operational set theory with unbounded existential quantification and power set. *Annals of Pure and Applied Logic*, 160, 2009.

[42] Gerhard Jäger. Operations, sets and classes. In Clark Glymour, Wei Wang, and Dag Westerstahl, editors, *Logic, Methodology, and Philosophy of Science: Proceedings of the Thirteenth International Congress*. College Publications, 2009.

[43] Gerhard Jäger, Reinhard Kahle, Anton Setzer, and Thomas Strahm. The proof-theoretic analysis of transfinitely iterated fixed point theories. *The Journal of Symbolic Logic*, 64:53–67, 1999.

[44] Gerhard Jäger and Jürg Krähenbühl. Σ^1_1 choice in a theory of sets and classes. In Ralf Schindler, editor, *Ways ofProof Theory*, pages 283–314. Ontos Verlag, Frankfurt, 2010.

[45] Reinhard Kahle. Truth in applicative theories. *Studia Logica*, 68:103–128, 2001.

[46] Reinhard Kahle. Universes over frege structures. *Annals of Pure and Applied Logic*, 119:191–223, 2003.

[47] Richard Kaye. *Models of Peano Arithmetic*. Clarendon Press, Oxford, 1991.

[48] Jeffey Ketland. Deflationism and tarski's paradise. *Mind*, 108:69–94, 1999.

[49] Henryk Kotlarski, Stanislaw Krajewski, and Alistair Lachlan. Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin*, 24:283–293, 1981.

[50] Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–716, 1975.

[51] Alistair Lachlan. Full satisfaction classes and recursive saturation. *Canadian Mathematical Bulletin*, 24:295–297, 1981.

[52] Graham Leigh and Michael Rathjen. An ordinal analysis for theories of self-referential truth. *Archive for Mathematical Logic*, 49:213–247, 2010.

[53] Hannes Leitgeb. What theories of truth should be like (but cannot be). *Philosophy Compass*, 2:276–290, 2007.

[54] Per Lindström. *Aspects of Incompleteness*. Lecture Notes in Logic 10. A K Peters, Ltd, Massachusetts, second edition, 2003.

[55] Grzegorz Malinowski. *Many-Valued Logics*. Oxford University Press, Oxford, 1993.

[56] Robert L. Martin and Peter W. Woodruff. On representing 'true-in-L' in L. *Philosophia*, 5:213–217, 1975.

[57] Vann McGee. How truthlike can a predicate be? *Journal of Philosophical Logic*, 14:399–410, 1985.

[58] Vann McGee. *Truth, Vagueness, and Paradox*. Hackett, Indianapolice, 1991.

[59] Vann McGee. Maximal consistent sets of instances of tarski's schema (t). *Journal of Philosophical Logic*, 21:235–241, 1992.

[60] Vann McGee. In praise of the free lunch: Why disquotationalists should embrace compositional semantics. In Thomad Bolander, Vincent F. Hendricks, and Stig A Pedersen, editors, *Self-Reference*, pages 95–120. CSLI Publications, Stanford, 2006.

[61] Sheard Michael. Truth, provability and naive criteria. In Volker Halbach and Leon Horsten, editors, *Principles of truth*, pages 169–181. Ontos Verlag, Frankfurt, 2002.

[62] Andrzej Mostowski. Some impredicative definitions in the axiomatic set-theory. *Fundamenta Mathematicae*, 37:87–110, 1950.

[63] Karl-Georg Niebergall. On the logic of reducibility: Axioms and examples. *Erkenntnis*, 53:27–61, 2000.

[64] Wolfram Pohlers. Subsystems of set theory and second order number theory. In Samuel Buss, editor, *Handbook of Proof Theory*, pages 209–336. Elsevier, Amsterdam, 1998.

[65] Wolfram Pohlers. Subsystems of set theory and second order number theory. In Samuel Buss, editor, *Handbook of Proof Theory*, pages 209–336. Elsevier, Amsterdam, 1998.

[66] Wolfram Pohlers. *Proof Theory*. Springer, Berlin, 2009.

[67] Graham Priest. Paraconsistent logic. In Dov Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 6, pages 287–393. Kluwer, Dordrecht, 2002.

[68] Michael Rathjen. The role of parameters in bar rule and bar induction. *The Journal of Symbolic Logic*, 56:715–730, 1991.

[69] Michael Rathjen. The realm of ordinal analysis. In S. Barry Cooper and John K. Truss, editors, *Sets and Proofs*, pages 219–279. Cambridge University Press, Cambridge, 1999.

[70] William N Reinhardt. Remarks on significance and meaningful applicability. In Luiz Paulo de Alcantara, editor, *Mathematical Logic and Formal Systems*, pages 227–242. Marcel Dekker, New York, 1985.

[71] William N Reinhardt. Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic*, 15:219–251, 1986.

[72] Greg Restall. *An Introduction to Substructural Logic*. Routledge, London, 2000.

[73] Kentaro Sato. The strength of extensionality II –weak weak set theories without infinity–. To appear in *Annals of Pure and Applied Logic*, 2010.

[74] Stewart Shapiro. Proof and truth: Through thick and thin. *The Journal of Philosophy*, 95:493–521, 1998.

[75] Moh Shaw-Kwei. Logical paradoxes for many-valued systems. *The Journal of Symbolic Logic*, 19:37–40, 1954.

[76] Michael Sheard. A guide to truth predicates in the modern era. *The Journal of Symbolic Logic*, 59:1032–1054, 1994.

[77] Stephen G. Simpson. *Subsystems of Second Order Arithmetic*. Cambridge University Press, Cambridge, 2009.

[78] Thomas Strahm. Autonomous fixed point progressions and fixed point transfinite recursion. In Samuel Buss, editor, *Logic Colloquium '98*, volume 13 of *ASL Lecture Notes in Logic*, pages 449–464. A K Peters, 2000.

[79] Gaisi Takeuti. *Proof Theory*. North-Holland, Amsterdam, second edition, 1987.

[80] Alfred Tarski. The semantic conception of truth and the foundation of semantics. *Philosophy and Phenomenological Research*, 4:341–376, 1944. Reprinted in: Simon Blackburn and Keith Simmons eds. *Truth*. Oxford University Press, Oxford, 1999: 115–143.

[81] Alfred Tarski. Truth and proof. *Scientific American*, 220:63–77, 1969.

[82] Alfred Tarski. The concept of truth in formalized languages. In John Corcoran, editor, *Logic, Semantics, Meta-mathematics*, pages 152–278. Hackett, Indianapolice, second edition, 1983.

[83] Alfred Tarski and Robert Vaught. Arithmetical extensions of relational systems. *Compositio Mathematica*, 13:81–102, 1956.

[84] Bas van Fraassen. Presupposition, implication, and self-reference. *The Journal of Philosopohy*, 65:136–152, 1968.

[85] Jouko Vänäänen. Second-order logic and foundation of mathematics. *The Bulletin of Symbolic Logic*, 7:504–520, 2001.

[86] Stephen Yablo. Paradox without self-reference. *Analysis*, 53:251–252, 1993.