Truth and Dependence

MSc Thesis *(Afstudeerscriptie)*

written by

Ruiting Jiang

under the supervision of **Dr. Thomas Schindler**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
June 27, 2025	Dr. Aybüke Özgün (Chair)
	Dr. Giorgio Sbardolini
	Dr. Luca Incurvati
	Dr. Thomas Schindler

Institute for Logic, Language and Computation

Abstract

Beginning with the study of the Liar paradox, philosophers have proposed several competing theories of truth, each built on different intuitions, and they provide distinct classifications of sentences as true, false, paradoxical, or hypodoxical. In the current literature, the two dominant approaches are Kripke's minimal fixed point construction and the Revision Theory of Truth. This thesis introduces a novel alternative: resting on a new underlying intuition that yields an alternative classification. I will argue that this theory is very natural and, in many respects, superior to existing accounts.

In our framework, every sentence corresponds to a function, which is determined by the sentences it depends on together with the T-schema. We then classify each sentence by the number of fixed points its associated function has. After presenting the formal theory, we compare our theory with Kripke's minimal fixed points and with the Revision Theory, showing how our approach admits certain circular tautologies without arbitrariness and captures a broader range of non-paradoxical puzzles. Finally, although the development takes place in an infinite propositional language, the last chapter sketches how these ideas can be adapted to first order logic, outlines the remaining obstacles, and suggests possible strategies for overcoming them.

Acknowledgements

I would like to express my gratitude to Thomas Schindler. I deeply appreciate all the time and guidance you have provided, not only for my master's thesis but also throughout my philosophical studies. Our meetings have always been productive and enlightening, and your insights have helped me approach problems more deeply, often clarifying the path forward. Beyond guiding me in my thesis work, your invaluable advice on improving my philosophical writing, selecting courses, and identifying essential papers and resources has been crucial. Coming from a purely mathematical background, I have learned so much from your supervision and suggestions, and I am now more confident and prepared to continue my journey in philosophy. I feel very lucky that I have studied under your guidance during my time at the ILLC.

I am also very grateful to Giorgio Sbardolini and Luca Incurvati for serving on my thesis committee and carefully reading my work. Your feedback during the defense was invaluable and will certainly inform my future research.

Lastly, I want to thank my family and friends for their support throughout my academic journey. I am particularly grateful to Minzhe Li, Tenyo Takahashi, and Fangjing Xiong for their thoughtful discussions on the topics of this thesis, which helped me clarify my ideas. I would also like to thank Jessica Shiqing Zhang for her encouragement and support throughout my time at the ILLC, which was invaluable to me and a great comfort during the challenging times of my studies. I am especially thankful to my parents, Caiyuan Jiang and Jing Liu, whose constant support has made it possible for me to pursue my studies in philosophy. Without their encouragement, I would not be where I am today, and I am grateful for everything they have done for me.

Contents

Chapter 1

Introduction

The Liar paradox has a long tradition in philosophy. Let the sentence L be "L is not true". Then we can find a contradiction: assume L is true, then what it says must be the case, so L is not true; assume L is not true, then, since this is exactly what L says, it is true. Therefore, L is true if and only if L is not true, a contradiction since a sentence must be either true or not true, but not both. Note that this is an issue for classical logic because one can conclude anything from a contradiction, and thus the logical system becomes trivial. Moreover, there is an important principle figuring in the above reasoning: Tarski's T-schema, which states that a sentence ϕ is true if and only if what it says is the case.

Tarski (1956) attempted to solve this problem by giving a hierarchy of truth, each applicable to sentences in a certain level. However, as pointed out by (Kripke, 1975), it suffers from a significant problem: failing to respect our use of "true" in natural language (Kripke, 1975). Two dominant hierarchy-free solutions have been proposed in the literature — Kripke's minimal fixed point (Kripke, 1975) and the Revision Theory of truth (Gupta and Belnap, 1993; Herzberger, 1982a). However, I will argue in later chapters that neither theory is entirely satisfactory.

The aim of this thesis is to advocate for another theory of truth which keeps "true" hierarchy-free. In this framework, each sentence corresponds to a function, and the patterns of fixed points of these functions are used to classify the semantic status of sentences. I will argue that this theory (1) is not an ad-hoc solution just for the Liar but instead give a general criterion for classifying the semantic status of sentences that refer to each other, (2) provides an explanation for the paradox by (2.1) acknowledging the paradoxical phenomenon in the Liar sentence (and other problematic sentences) and (2.2) explaining it in a way that is harmonious with our intuition, and (3) it has certain advantage over some dominant hierarchy-free theories of truth — namely, Kripke's minimal

The introduction part of this Chapter and Section 1.1 is based on an essay I wrote for an individual project done with Dr. Thomas Schindler.

fixed points and the Revision Theory. Moreover, the theory has robust applications in the study of paradoxes.

The structure of the thesis is organised as follows. In the remainder of the first chapter, I present Tarski's hierarchical solution of the Liar, and I use Kripke's argument to show that the truth predicate it gives does not respect the use of "true" in natural language. Then I will motivate the proposed theory of truth by reflecting on how we reason with paradoxes (and more generally, sentences that refer to each other) pre-theoretically, and I will introduce the idea of corresponding a function to each sentence.

In Chapter 2, I present the formal theory of truth in a language of infinite propositional logic. I will present the language and the denotation function, and how to model our sentences of interest using this language. I will also define the dependence relation between sentences in this language, and I will show how to assign a function to each sentence that takes as input a hypothetical truth value and outputs the truth value according to Tarski's T-schema. In the last two sections I will discuss the issue with the interpretation of the theory. These two sections are marked with a star, and skipping them in a first reading will not affect the understanding of the rest of this thesis.

In the next two Chapters (3 and 4), I compare the proposed theory with two dominant theories of truth — Kripke's minimal fixed points and the Revision Theory of truth. Since they are originally developed in the first order language, instead of the propositional language I use, I will first present basic ideas of both theories and then discuss how the underlying idea can be used to yield a theory in the propositional language. Then I will formally compare the differences between the two theories and my proposed theory.

In Chapter 5, we will see how the theory can be applied to answer several questions about paradoxes, and how it can be adapted to meet certain intuitions. In particular, the theory suggests a sense in which the hypodoxical sentences, like the Truth Teller, is also paradoxical; it can formalise our intuition that the Liar circle can be reduced to the Liar paradox; and, if one prefers a theory behaving more classically in some aspects, the theory can be modified to respect these intuitions.

Lastly, in Chapter 6, I will sketch how the theory could be developed in a first order language and discuss the remaining challenges. Moreover, I will propose two possible strategies to overcome these difficulties, and we will see that each strategy leads to research questions that are interesting in their own right.

1.1 Hierarchical Solutions to the Liar Paradox

In this section, I briefly discuss Tarski's solution to the paradox by employing a hierarchy of truth predicates. I will argue that the solution fails to resolve the paradox, as it is a paradox in natural language.

The central idea of Tarski's theory is that truth of sentences in one language can

only be talked about in a richer language. Starting from a language $\mathcal{L}_0$, Tarski observes that it is impossible — due to the Liar paradox — for $\mathcal{L}_0$ to have a predicate $True$ whose extension is exactly the set of all true sentences in $\mathcal{L}_0$. Therefore, we need another language $\mathcal{L}_1$ that extends $\mathcal{L}_0$ by adding a predicate $True_0$ to talk about true sentences in $\mathcal{L}_0$. However, the same observation applies to $\mathcal{L}_1$, so we need another language $\mathcal{L}_2$ extending $\mathcal{L}_1$ by yet another predicate $True_1$ to talk about true sentences in $\mathcal{L}_1$. This process goes on indefinitely, and we obtain a linear hierarchy of languages $\mathcal{L}_0$, $\mathcal{L}_1$, $\mathcal{L}_2$, ..., each containing a truth predicate applicable only to sentences from lower levels.

According to this theory, the Liar sentence in one language, for example, $\gamma = \neg True_n(\ulcorner\gamma\urcorner)$, is indeed not in the extension of $True_n$, so it is true[1]. However, this truth cannot be expressed in the language $\mathcal{L}_n$ where γ is constructed, but only in higher languages, so we have $True_{n+1}(\ulcorner\gamma\urcorner)$, instead of $True_n(\ulcorner\gamma\urcorner)$, and hence there is no contradiction.

Nonetheless, this theory does not fit well with our practice of natural language. In natural language, we do use the predicate "true" to talk about truth of sentences in the same language. Moreover, as pointed out by Kripke (1975), the theory cannot make sense of some sentences that are perfectly sensible in natural language. He gives the following example: suppose Dean says "everything Nixon says about Watergate is not true" while Nixon says "everything Dean says about Watergate is not true". Intuitively, if Dean said something true about Watergate, then Nixon's sentence is false, and if Nixon's other utterances about Watergate are all false, then Dean's sentence is true.

However, according to Tarski's theory, these two sentences cannot be grammatical — the truth predicate used by Dean must lie in a higher level in the hierarchy of language than all utterances by Nixon about Watergate, so in particular it must lie in a higher level than the truth predicate in Nixon's sentence above. However, the same argument applies to Nixon's truth predicate, so we have a violation of the linear hierarchy of languages.

Therefore, by giving a hierarchy of languages, Tarski's theory diverges too much from natural language, and hence fails to resolve the paradox as it is presented in natural language.

We need a theory of truth that keep the truth predicate non-hierarchical and applicable to sentences in the same language. In the current literature, Kripke's theory and the Revision Theory of truth are two dominant hierarchy-free theories of truth. However, as I will argue in Chapter 3 and Chapter 4, both theories are not satisfactory. Hence, I will propose a new theory of truth in this thesis.

[1]This is because the extension of $True_n$ consists precisely of the set of true sentences in the language $\mathcal{L}_{n-1}$. As γ contains the predicate $True_n$ — which is not a symbol in $\mathcal{L}_{n-1}$ — γ is not a sentence in $\mathcal{L}_{n-1}$, and therefore does not belong to the extension of $True_n$. Thus, $\gamma = \neg True_n(\ulcorner\gamma\urcorner)$ is true.

1.2 A Theory of Truth via Functions

Let us reflect on how we reason with paradox — and more generally, a collection of sentences referring to the truth of each other — pre-theoretically. Consider the Liar paradox "this sentence is not true." Why do we think this is a paradox? Or, by what procedure do we decide that this is a paradox? We do not simply say that this sentence refers to itself, so it must be problematic, but we reason with Tarski's T-schema: assume it is true, then by T-schema it is false, and if it is false, then by T-schema it is true. However, a sentence cannot be both true and false, so we conclude that the sentence is paradoxical.

Consider another example, the Truth Teller sentence "this sentence is true." Why do we think this is problematic? We reason again: assume it is true, then by T-schema it is true, and if it is false, then by T-schema it is false. Then we cannot decide which value to ascribe to the sentence, but we should not need any extra information to decide its truth value — the sentence only says something about itself. Hence, we conclude that the sentence is problematic.

Lastly, consider the typical logic puzzles of Knights and Knaves (Smullyan, 1978): people from a village are either knaves (who always tell lies) or knights (who always tell the truth), and they each say something about the identity of other villagers. For example, suppose we encounter two villagers, A and B. A says "Either B is knight and I am a knave, or B is a knave and I am a knight", while B says "We are both knights or both knaves" (CSCI E-80, 2021). How can we decide the identities of A and B? We test all four possibilities:

1. Both knights. Then A is lying — a contradiction to our assumption that A is a knight.
2. Both knaves. Then B is telling the truth — a contradiction.
3. A knave, B knight. Then B is lying — a contradiction.
4. A knight, B knave. Then A tells the truth and B lies — this is consistent with our assumption.

Thus, the only coherent solution is that A is a knight and B is a knave, so we conclude that this is the answer.

Why do we think that this is a puzzle and have fun solving it instead of claiming it a paradox? This is because we try and find a truth configuration for each sentence such that under and only under this configuration all the statements are coherent.

In all these examples, what we do is to give each relevant[2] sentence a hypothetical truth value, and reason, using T-schema, to obtain another truth value for every sentence

[2]Note that we only give hypothetical truth values to *relevant* sentences – take the example of Liar paradox, to decide it is paradoxical, we only give hypothetical truth value for the Liar, but not other

involved. If we find that the resulting truth value is the same as the hypothetical truth value we begin with, and if this is the only configuration of truth assignment which gives back the original values, then we say that there is no paradox, and we declare that we have solved a logic puzzle! If no such configuration exists, then we say that there is a paradox, and if there are multiple such configurations, then we say that the sentences given are problematic.

A natural way to model our reasoning here is to let each sentence correspond to a function, which takes as input a hypothetical truth value for each relevant sentence, and outputs their truth values according to the T-schema. We then check whether the function has a fixed point. Only when it has a unique fixed point, do we say that the corresponding sentence is not problematic.

For example, the Liar sentence corresponds to a function f such that $f(1) = 0$ and $f(0) = 1$. The function has no fixed point, so the Liar sentence is paradoxical. The Truth Teller sentence corresponds to a function f such that $f(1) = 1$ and $f(0) = 0$. The function has two fixed points, so the Truth Teller sentence is problematic. The logic puzzle corresponds to a multivalued function, taking as input a hypothetical truth value for each sentence and giving an output truth value for each sentence according to what they are saying[3]. If the function has a unique fixed point, then the fixed point gives the answer to the puzzle[4]; otherwise, we complain that the designer of the puzzle has made a mistake — either by making the puzzle inconsistent (when there is no fixed point) or by giving insufficient information (when there are multiple fixed points).

According to the idea above, Tarski's T-schema should be understood as a rule to assign truth to each sentence. Moreover, just as assigning a value to a variable in mathematics, the assignment can fail when the rule gives no unique value: for example, if I assign a value to x according to the rule "x is the number that satisfies $x^2 = 4$", then the assignment fails because there are two numbers that satisfy the equation. This assignment does not directly give a value to x but instead states a condition that x needs to satisfy, and thus it fails when the condition does not pick out a unique value.

Similarly, when T-schema is applied to a sentence like the Liar, it does not assign a truth value to the sentence but instead states a condition that the value needs to satisfy. When the condition cannot be met or there are multiple ways to meet the condition, then the assignment rule fails. We would only call a sentence true if the assignment rule does

irrelevant sentences like the Truth Teller. Similarly, in the Knights-Knaves puzzle, we consider only the four possible identity assignments for A and B, ignoring any other villagers which might also exist.

[3]Of course, when solving actual puzzles, there are techniques where you can start with the most "suspicious" people, but in general this is the underlying idea.

[4]Our previous example corresponds to a function $f : \{0,1\}^2 \to \{0,1\}^2$ such that $f(1,1) = (0,1)$ — which means that if we assume both are knights (1) then, according to what they say, A is a knave (0) and B is a knight (1) — $f(0,0) = (0,1)$, $f(0,1) = (1,0)$, and $f(1,0) = (1,0)$. The function has a unique fixed point $(1,0)$, which means that A is a knight and B is a knave. The formulation in Section 2.2 will be slightly different due to some technicalities, but this is the essential idea.

uniquely assign the value 1 to this sentence.

This is the idea I want to explore in this thesis.

Chapter 2

A Theory of Truth in The Language of Infinite Propositional Logic

In this chapter, I formulate the proposed theory of truth using a language of infinite propositional logic developed in (Rabern, Rabern, and Macauley, 2012). I will first present the formal language. Then, I will develop the idea of corresponding a function to each sentence in the language, and use these functions to classify the semantic status of sentences in the language. After the classification, I will discuss several properties of the set of "true" sentences, and I will show that the classification is independent of some irrelevant choices within the language. The last two sections will be devoted to the discussion of the limitation and the interpretation of the theory. These two sections are marked with a star, and they can be skipped in a first reading without affecting the understanding of the rest of this thesis.

2.1 The Language of Infinite Propositional Logic

In this section, we will describe the formal language and see how it can be applied to model semantic paradoxes like the Liar.

We work in an infinite propositional language $\mathcal{L}_S$ developed in (Rabern, Rabern, and Macauley, 2012), which has a set S of propositional variables. Through the use of a denotation function, these variables will be interpreted as names of sentences in this language.

The choice of this language is because it provides an especially intuitive framework for representing the semantic paradoxes. Although it contains no predicate symbols and is propositional, combining it with the denotation function gives us all the machinery we need to analyse important examples in the study of semantic paradoxes — such as the Liar, the Truth Teller, and Yablo's paradox[1].

[1]Introduced by Yablo in (1993) as a paradox without self-reference.

Definition 2.1 (The Language $\mathcal{L}_S$)**.** *A sentence in the language $\mathcal{L}_S$ is defined recursively as follows:*

1. *$\bot$, $\top$ are sentences ($\bot, \top \notin S$);*
2. *any $s \in S$ is a sentence;*
3. *if ϕ and ψ are sentences, then so are $\neg\phi$ and $\phi \wedge \psi$;*
4. *for any set of sentences Φ, $\bigwedge \Phi$ is a sentence.*
5. *Nothing else is a sentence.*

Let us denote the set of all sentences in $\mathcal{L}_S$ as $Sent(\mathcal{L}_S)$. The symbols $\rightarrow$ and $\vee$ are defined in the standard way, and I will use $\bigvee \Phi$ as the abbreviation for $\neg \bigwedge \neg\Phi$, where $\neg\Phi = \{\neg\phi \mid \phi \in \Phi\}$.

We move on to define how to evaluate the truth value of a sentence in this language. As in classical propositional logic, we start by giving each propositional letter a truth value.

Definition 2.2 (Valuation)**.** *A* valuation *v is a function from S to $\{1, 0\}$.*

Then this valuation is extended recursively to all sentences in the language $\mathcal{L}_S$:

Definition 2.3 (Extended Valuation)**.** *Any valuation v can be extended to a function $\bar{v}$ from $Sent(\mathcal{L}_S)$ to $\{1, 0\}$ as follows:*

1. *$\bar{v}(\bot) = 0$ and $\bar{v}(\top) = 1$;*
2. *$\bar{v}(s) = v(s)$ for any $s \in S$;*
3. *$\bar{v}(\neg\phi) = 1 - \bar{v}(\phi)$;*
4. *$\bar{v}(\phi \wedge \psi) = \min\{\bar{v}(\phi), \bar{v}(\psi)\}$;*
5. *$\bar{v}(\bigwedge \Phi) = \min\{\bar{v}(\phi) \mid \phi \in \Phi\}$.*

We identify $\bar{v}$ with v when the context is clear, and we denote $v(\phi)$ as $[\![\phi]\!]_v$.

Note that the truth value of a sentence ϕ under some valuation v only depends on what values v assigns to the propositional letters in ϕ.

Lemma 2.4. *For any $\phi \in Sent(\mathcal{L}_S)$, and any valuation v, v', if $v(s) = v'(s)$ for all propositional letters s occurring in ϕ, then $[\![\phi]\!]_v = [\![\phi]\!]_{v'}$.*

Proof. This is straightforward by induction on the complexity of ϕ. □

Therefore, we can just specify what v assigns to the propositional letters in ϕ instead of the entire valuation v when we are only interested in ϕ.

Definition 2.5 (Satisfaction)**.** *We say that a valuation* v satisfies *a sentence* ϕ *if* $[\![\phi]\!]_v = 1$*, and we denote it as* $v \models \phi$*.*

We now show how to represent paradoxes in this language. The idea is that the propositional variables in S will be used as names of sentences. They are sentences themselves, while they can also be used to refer to other sentences. This is a natural move in light of the semantic paradoxes. For example, consider the Liar sentence $L =$ "L is not true". L is a sentence, but it is also used as a name to refer to the sentence "L is not true", which contains L as a sub-sentence.

Definition 2.6 (Denotation Function)**.** *Given a set of propositional variables* S*, a* denotation function $d : S \to Sent(\mathcal{L}_S)$ *is a function that assigns each propositional variable a sentence in the language.*

Let us see how the Liar sentence and the Truth Teller sentence can be represented in this language:

Example 2.7. *Let* $s_1, s_2 \in S$ *and* $d(s_1) = \neg s_1$*,* $d(s_2) = s_2$*. Then* s_1 *models the Liar sentence and* s_2 *models the Truth Teller sentence.*

Thus, $s \in S$ is also the name of the sentence $d(s) \in Sent(\mathcal{L}_S)$. Let us follow (Rabern, Rabern, and Macauley, 2012) and call (S, d) a sentence system.

2.2 Dependence, Ascriptions, and Truth

Now, we have enough devices to develop the idea of corresponding a function to each sentence. Recall the motivating examples in Section 1.2. Given any sentence, we will give a hypothetical truth value to all sentences that are relevant to the sentence. Besides the sentence itself, we need to find out sentences that the given sentence depends on. Therefore, we first need to define this dependence relation. In the propositional language we use, there is a straightforward way to make this definition: a sentence s depends on another sentence t if $d(s)$ contains t as a subformula. Moreover, we require that this relation is transitive, so that if s depends on t and t depends on u, then we will also say that s depends on u.

Definition 2.8 (Dependence Relation)**.** *We first define a relation* $R \subseteq S \times S$ *on the set of names of sentences as follows:* sRt *if and only if* $d(s)$ *contains* t *as a subformula. Let* R^* *be the transitive closure of* R*. We say* s depends *on* t *if* sR^*t*.*

Now we collect all sentences that s depends on.

Definition 2.9 (Dependence Set)**.** *We define the* dependence set D_s *for* s *as the set of sentences that* s *depends on:* $D_s = \{t \in S \mid sR^*t\}$*.*

As in our motivating examples, given a collection of sentences which depend on each other in a certain way, we can give each sentence a hypothetical truth value and reason according to what they say to obtain another truth value for each of them. For any sentence s the collection of sentences that are relevant in this process is exactly the sentences in D_s, plus s itself — since regardless of whether s depends on itself, we give a hypothetical truth value to all the sentences in this collection. As this collection of sentences will be used to determine the number of variables that the function corresponding to s will take, let us call it the *variable set* of s.

We will see latter[2] that we do not need to give a hypothetical truth value to s itself when s does not depend on itself. However, at this stage, I believe it is best to stick with the intuition as close as possible. The fact that we do not need to include s is best treated as a discovery, or — for those with the intuition that we might not need to give a hypothetical value for s even informally — at least a confirmation that this intuition is formally correct.

Definition 2.10 (Variable Set)**.** *We define the* variable set V_s *for* s *as the set of sentences that* s *depends on plus* s *itself:* $V_s = D_s \cup \{s\}$.

Example 2.11. *1. Let* $s \in S$ *and* $d(s) = \neg s$. *Then* $D_s = V_s = \{s\}$.

2. Let $s_1, s_2 \in S$ *and* $d(s_1) = s_2$ *and* $d(s_2) = \top$. *Then* $D_{s_1} = \{s_2\}$ *and* $V_{s_1} = \{s_1, s_2\}$, *while* $D_{s_2} = \emptyset$ *and* $V_{s_2} = \{s_2\}$. *Note that* $\top \notin D_{s_2}$ *because by definition* $D_s \subseteq S$ *for all* $s \in S$, *while* $\top$ *is not in* S.

By a slight abuse of language, we identify V_s with a list ordered by the ordinals. There is no deep philosophical reason for doing this, but it will help us pick out some $t \in V_s$ by its index in the list in our formal details. Thus, intuitively, one can still think of V_s simply as the collection of sentences that are relevant to s, and now we are able to identify the location of each sentence in this collection.

Formally, we will write $V_s(\alpha)$ for the α-th coordinate of the list V_s. Moreover, for technical convenience, we stipulate that $V_s(0) = s$. That is, we assume that the sentence s itself is always put in the first coordinate of this list. Also, we say that the index of a sentence $t \in V_s$ is the unique ordinal α_t such that $V_s(\alpha_t) = t$. For notational simplicity, for $\bar{x} \in \{0,1\}^\kappa$, where κ is some cardinal, we often write x_i for $\bar{x}(i)$, the i-th coordinate of $\bar{x}$.

Each sentence s corresponds to a function $f_s : \{0,1\}^{|V_s|} \to \{0,1\}^{|V_s|}$ that takes as input a hypothetical truth value for each sentence in V_s and outputs their truth value according to the hypothetical values and the denotation function:

Definition 2.12 (Ascription Function)**.** *Let* $s \in S$ *be a sentence in the sentence system* (S, d). *We define the* ascription function $f_s : \{0,1\}^{|V_s|} \to \{0,1\}^{|V_s|}$ *coordinate-wise as*

[2]This will follow from Lemma 2.31, which says that as long as all sentences that s depends on are either true or false, then s itself is either true or false (so that it cannot be paradoxical, or problematic in some other sense.)

follows: $f_s(\bar{x})(\alpha) = [\![d(V_s(\alpha))]\!]_v$*, where the valuation* v *is such that for a sentence* $t \in V_s$*, we have* $v(t) = x_i$ *for* $t = V_s(i)$*.*

Despite the heavy notation, the idea underlying this definition is very simple. An element $\bar{x} \in \{0,1\}^{|V_s|}$ is interpreted as giving a hypothetical truth value to each sentence in V_s: the i-th coordinate of $\bar{x}$ gives the hypothetical truth value of the i-th sentence in V_s. Then $\bar{x}$ induces a valuation v such that for any $t = V_s(i)$, we have $v(t) = x_i$. We then use this v to calculate the truth value for each sentence $t \in V_s$ according to $d(t)$, and then collect the resulting truth values as $f_s(\bar{x}) \in \{0,1\}^{|V_s|}$. In particular, the α-th coordinate of $f_s(\bar{x})$ — which is what we denoted by $f_s(\bar{x})(\alpha)$ — gives the truth value of the α-th sentence in V_s according to the valuation v induced by $\bar{x}$, which is denoted by $[\![d(V_s(\alpha))]\!]_v$. Let us see how this works in some typical examples.

Example 2.13. *1. The Liar: let* $s \in S$ *be such that* $d(s) = \neg s$*. Then* $V_s = \{s\}$ *and* f_s *is a function from* $\{0,1\}$ *to* $\{0,1\}$*. We have* $f_s(0) = [\![d(s)]\!]_v$ *where* $v(s) = 0$*. Thus,* $f_s(0) = [\![d(s)]\!]_v = [\![\neg s]\!]_v = 1$*. This means that if we give the Liar a hypothetical truth value* 0*, then, according to what it says, we conclude that it is in fact true, so we obtain a new truth value of* 1*. Similarly,* $f_s(1) = [\![d(s)]\!]_v = [\![\neg s]\!]_v = 0$*, where* $v(s) = 1$*. This means that if we give the Liar a hypothetical truth value* 1*, then, according to what it says, we conclude that it is in fact false. Thus, we have* $f_s(x) = 1 - x$*. That is, the Liar corresponds to a function that always flips the input truth value.*

2. The Truth Teller: let $s \in S$ *be such that* $d(s) = s$*. We similarly have* $f_s : \{0,1\} \to \{0,1\}$*. This time,* $f_s(0) = [\![d(s)]\!]_v = [\![s]\!]_v = 0$*, where* $v(s) = 0$*; and* $f_s(1) = [\![d(s)]\!]_v = [\![s]\!]_v = 1$*, where* $v(s) = 1$*. Thus, we have* $f_s(x) = x$*. That is, the Truth Teller corresponds to a function that always returns the input truth value as it is.*

3. We now give an intuitively unproblematic case. Consider a sentence L_1 *which says "*L_2 *is true", while* L_2 *just says some obvious truth, for example "*$1 = 1$*". Clearly, we should conclude that both* L_1 *and* L_2 *are true.*

This can be modelled in the formal language as follows. Let $s_1, s_2 \in S$ *where* $d(s_1) = s_2$ *and* $d(s_2) = \top$*. We have seen from the previous example that* $V_{s_1} = \{s_1, s_2\}$ *and* $V_{s_2} = \{s_2\}$*. Therefore, we have* $f_{s_1} : \{0,1\}^2 \to \{0,1\}^2$*.*

Given an input, say, $(0,0)$*, which induces a valuation* v *such that* $v(s_1) = v(s_2) = 0$*. We calculate* $f_{s_1}(0,0) \in \{0,1\}^2$ *coordinate-wise. The* 0*-th coordinate* $f_{s_1}(0,0)(0)$ *is given by* $[\![d(V_s(0))]\!]_v = [\![d(s_1)]\!]_v = [\![s_2]\!]_v = 0$*; and the* 1*-th coordinate* $f_{s_1}(0,0)(1)$ *is given by* $[\![d(V_s(1))]\!]_v = [\![d(s_2)]\!]_v = [\![\top]\!]_v = 1$*. Thus, we have* $f_{s_1}(0,0) = (0,1)$*.*

This is indeed how our intuitive reasoning works: if we assume both L_1 *and* L_2 *are false, then according to what they say,* L_1 *is false and* L_1 *is true.*

Similarly, we can calculate $f_{s_1}(0,1) = (1,1)$, $f_{s_1}(1,0) = (0,1)$, *and* $f_{s_1}(1,1) = (1,1)$, *i.e.* $f_{s_1}(x,y) = (y,1)$.

On the other hand, we have $f_{s_2} : \{0,1\} \to \{0,1\}$, *and* $f_{s_2}(0) = [\![d(s_2)]\!]_v = [\![\top]\!]_v = 1$ *and* $f_{s_2}(1) = [\![d(s_2)]\!]_v = [\![\top]\!]_v = 1$, *i.e.* $f_{s_2}(x) = 1$.

4. *Liar circle: let* $s_0, s_1, \ldots, s_n \in S$ *be such that* $d(s_i) = s_{i+1}$ *for* $i < n$, *and* $d(s_n) = \neg s_0$. *Then* $V_{s_0} = \{s_0, s_1, \ldots, s_n\}$ *and* $f_{s_0} : \{0,1\}^{n+1} \to \{0,1\}^{n+1}$. *One can check that* $f_{s_0}(\bar{x}) = (x_1, x_2, \ldots, x_n, 1 - x_0)$.

5. *Truth Teller circle: let* $s_0, s_1, \ldots, s_n \in S$ *be such that* $d(s_i) = s_{i+1}$ *for* $i < n$, *and* $d(s_n) = s_0$. *Then* $V_{s_0} = \{s_0, s_1, \ldots, s_n\}$ *and* $f_{s_0} : \{0,1\}^{n+1} \to \{0,1\}^{n+1}$. *One can check that* $f_{s_0}(\bar{x}) = (x_1, x_2, \ldots, x_n, x_0)$.

6. *Yablo's paradox: let* $\{s_i \mid i \in \omega\} \subseteq S$ *be such that* $d(s_i) = \bigwedge_{i<j} \neg s_j$ *for* $i \in \omega$. *Then* $V_{s_0} = \{s_i \mid i \in \omega\}$ *and* $f_{s_0} : \{0,1\}^\omega \to \{0,1\}^\omega$. *One can check that the* α*-th coordinate of* $f_{s_0}(\bar{x})$ *is given by* $f_{s_0}(\bar{x})(\alpha) = min\{1 - x_i \mid \alpha \leq i\}$.

One may worry that for any sentence s, V_s is a set of sentences, but to obtain the function f_s we need to first order V_s into a list, and then the function might depend on how we order the sentences in V_s. For example, in the Liar circle, we have $V_{s_0} = \{s_0, s_1, \ldots, s_n\}$. We may use the list $\langle s_0, s_1, s_2 \ldots, s_n\rangle$ to obtain the function $f_{s_0}(\bar{x}) = (x_1, x_2, x_3, \ldots, x_n, 1 - x_0)$, or we may use $\langle s_0, s_2, s_1, \ldots, s_n\rangle$ to obtain a function $g_{s_0}(\bar{x}) = (x_1, x_3, x_2, \ldots, x_n, 1 - x_0)$. These two functions are clearly different. Therefore, indeed, the functions that a sentence corresponds to depend on how one orders its variable set.

However, for the study of paradox, these differences do not matter. This is because we are interested in whether a coherent truth assignment can be found for all the sentences in the variable set of a sentence, which is reflected not by the function itself but by the patterns of fixed points of the function, which we define below.

Definition 2.14 (Classisfication of Ascription Functions)**.** *Let* $s \in S$ *be a sentence in the sentence system* (S, d). *We say that the ascription* f_s *is*

1. **successful** *if the equation* $\bar{x} = f_s(\bar{x})$ *has a unique solution;*

2. **paradoxical** *if the equation* $\bar{x} = f_s(\bar{x})$ *has no solution;*

3. **hypodoxical** *if the equation* $\bar{x} = f_s(\bar{x})$ *has multiple solutions;*

We will explain the intuition behind this definition, but we first show that the order of the sentences in the variable set does not affect whether the ascription function is successful, paradoxical, or hypodoxical.

Lemma 2.15. *Let $s \in S$ be a sentence in the sentence system (S, d), and V_s be the variable set of s. Let V_s^1 and V_s^2 be two lists of sentences that contain the same sentences as V_s. Let f_s be the function corresponding to s with respect to V_s^1, and let g_s be the function corresponding to s with respect to V_s^2. Then f_s is successful (paradoxical, hypodoxical) if and only if g_s is successful (paradoxical, hypodoxical).*

Proof. Let $|V_s| = n+1$ and $\sigma : n+2 \to n+2$ be a permutation such that $V_s^2(i) = V_s^1(\sigma(i))$. Then $\bar{x} = (x_0, x_1, \ldots, x_n)$ is a fixed point of f_s if and only if $\sigma(\bar{x}) = (x_{\sigma(0)}, x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ is a fixed point of g_s. □

The intuition behind Definition 2.14 is clear. In a successful ascription, there is one and only one coherent truth assignment for all the sentences in the variable set of a sentence; in a paradoxical ascription, there is no coherent truth assignment for them; and in a hypodoxical ascription, there are multiple.

A tempting next step is to say that a sentence s is "true" if f_s is successful and the truth value of s is 1 under this unique coherent truth assignment — that is to say, to call s "true" if the fixed point $\bar{x}$ of f_s satisfies $x_0 = 1$. However, another worry arises. In the fixed point, not only is the truth value of s determined, but it simultaneously gives the truth value for all the sentences in the variable set of s — as we said, the unique fixed point is the coherent truth assignment for all those sentences. This is because the calculation shows that only when the sentences involved are together assigned the values according to the fixed point, the whole assignment is coherent. However, if s depends on t, then t also induces an ascription function f_t. This f_t might disagree with f_s with respect to the fixed point — for example, f_s might say there is a unique fixed point where t is assigned value 1, while f_t might say that there are many more fixed points. We give some examples of this situation.

Example 2.16.

1. *Let $s_1, s_2 \in S$, $d(s_1) = (s_1 \wedge s_2) \vee (\neg s_1 \wedge s_2) \vee (\neg s_1 \wedge \neg s_2)$ and $d(s_2) = s_2$. Then $V_{s_1} = \{s_1, s_2\}$ and $V_{s_2} = \{s_2\}$. One can check that $(1, 1)$ is the only fixed point of f_{s_1}. This means that only if both s_1 and s_2 are true do we have a coherent truth assignment for both s_1 and s_2. However, s_2 itself — which is just the Truth Teller — does not depend on s_1, and f_{s_2} does not have a unique fixed point.*

2. *Let $s_3, s_2 \in S$, $d(s_3) = \neg s_3 \vee s_2$ and $d(s_2) = s_2$. Then $V_{s_3} = \{s_3, s_2\}$ and $V_{s_2} = \{s_2\}$. One can check that $(0, 0)$ is the only fixed point of f_{s_3}, while f_{s_2} does not have a unique fixed point.*

If we just naively use the fixed point to determine the truth of sentences, we will face a contradiction. In both of the above examples, s_2 is the Truth Teller. According to s_1, both s_1 and s_2 are true, while according to s_3, both s_3 and s_2 are false. At the same time, s_2 itself says it does not have a definite truth value! This is a serious problem.

Fortunately, the problem is not with the theory, but with how one interprets the fixed point of the ascription functions. Let us focus on the second example from above, where we have $(0, 0)$ as the unique fixed point of f_{s_3}. This means that we can have a coherent truth assignment for both s_2 and s_3 only when both s_2 and s_3 are false. However, this does not mean that s_2 is confirmed to be false. It only means that if s_2 is false, then we have s_3 is false. In fact, modulo classical logic, this is exactly what s_3 says — $d(s_3) = \neg s_3 \vee s_2 \equiv s_3 \rightarrow s_2 \equiv \neg s_2 \rightarrow \neg s_3$. Then, when we do not have s_2 is false — the Truth Teller does not have a truth value due to indeterminacy — clearly we cannot conclude that s_3 is also false.

Therefore, to actually determine the truth status of a sentence s, it is not enough that f_s has a unique fixed point; we also need that for any sentence t such that s depends on t, f_t has a unique fixed point, and the truth value assignments given by the fixed point of f_t should agree with the truth value assignments given by the fixed point of f_s. We now make this formal.

Definition 2.17 (Naive Truth Value)**.** *Let $s \in S$ be a sentence in the sentence system (S, d) where f_s is successful. Let $\bar{x}$ be the fixed point of f_s. We say that the* naive truth value *of $t \in V_s$ is $x(\alpha_t)$, where α_t is the index of t in V_s.*

That is, the naive truth value of a sentence $t \in V_s$ is the truth value of t determined by the fixed point of the function f_s corresponding to s. Two ascription functions agree on a sentence t if they give the same naive truth value to t.

Definition 2.18 (Agreement)**.** *Let $s \in S$ be a sentence in the sentence system (S, d) and let $t \in D_s$, where f_s and f_t are both successful. We say that the ascription function f_s* agrees *with the ascription function f_t if the solution of the equation $\bar{x} = f_s(\bar{x})$ agrees with the solution of the equation $\bar{y} = f_t(\bar{y})$ in the sense that for all $t' \in V_t$, the naive truth value of t' as determined by f_s is the same as the naive truth value of t' as determined by f_t.*

We now formalise the condition where the truth value of a sentence can really be determined.

Definition 2.19 (Hereditarily Successful Ascription)**.** *Let $s \in S$ be a sentence in the sentence system (S, d). We say that the ascription f_s is* **hereditarily successful** *if it is:*

1. *successful;*
2. *for any $t \in V_s$, f_t is successful; and*
3. *f_s agrees with f_t for any $t \in V_s$.*

Notice that in the examples we gave above, we have cases where a function is successful, but it fails to be hereditarily successful because it violates the second condition. One might

wonder whether there are cases where both the first and the second condition are satisfied, but the third condition is violated — so that each ascription function gives some naive semantic truth assignments, but these assignments are contradictory. In these situations it might be unclear what to say about the sentences involved. However, fortunately, we can show that this cannot happen, and hence the definition of hereditary success could be simplified.

Lemma 2.20. *Let $s \in S$ be a sentence in the sentence system (S, d) and $t \in V_s$. If f_s and f_t are both successful, then f_s agrees with f_t.*

Proof. Let $\bar{x}$ be the unique fixed point of f_s. Let $\bar{y}$ be such that $y(\alpha_p) = x(\alpha_p)$ for all $p \in V_t$. Then $\bar{y}$ is a fixed point of f_t. Since f_t is successful, this is fixed point is unique. Therefore, f_s agrees with f_t. □

Therefore:

Corollary 2.21. *Let $s \in S$ be a sentence in the sentence system (S, d). The ascription f_s is hereditarily successful if and only if for any $t \in V_s$, f_t is successful.*

In fact, for any sentence s, we only need to check the sentences it depends on to determine whether s is hereditarily successful.

Lemma 2.22. *Let $s \in S$ be a sentence in the sentence system (S, d). The ascription f_s is hereditarily successful if and only if for any $t \in D_s$, f_t is successful.*

Proof. $\Rightarrow$: This is implied by the previous lemma.

$\Leftarrow$: By the previous lemma, it suffices to prove f_s is successful. If $s \in D_s$ then we are done. Otherwise, denote the i-th sentence in D_s as t_i. Let x_i be the naive truth value of t_i as determined by f_{t_i}, which is well-defined since f_{t_i} is successful. Let $x_0 = [\![d(s)]\!]_v$, where $v(t_i) = x_i$ for all i. Then $f_s(\bar{x}) = \bar{x}$, and $\bar{x}$ is the unique fixed point of f_s. □

Note a helpful technique we used several times in the above proofs:

Lemma 2.23. *Let $s \in S$ be a sentence in the sentence system (S, d). If f_s has a fixed point and $t \in V_s$, then f_t also has a fixed point.*

Collecting all the above, we also have the following useful corollary:

Corollary 2.24. *Let $s \in S$ be a sentence in the sentence system (S, d). The ascription f_s is hereditarily successful if and only if for all $t \in D_s$, f_t is hereditarily successful.*

Proof. $\Rightarrow$: This follows from the transitivity of the dependence relation.

$\Leftarrow$: this is implied by Lemma 2.22. □

Remark 2.25. *Note that, therefore, in the definition of the ascription function, we could have replaced V_s by D_s. By the corollary above, this will give an equivalent theory. However, we will keep using V_s, because it is more intuitive to think of all sentences in V_s as relevant when we reason about s.*

Finally, we can determine the semantic status of a sentence.

Definition 2.26 (Classisfication of Sentences)**.** *Let $s \in S$ be a sentence in the sentence system (S, d). We say that s is:*

1. **paradoxical** *if f_s is paradoxical;*
2. **hypodoxical** *if f_s is hypodoxical;*
3. **true** *if f_s is hereditarily successful and the solution of the equation $\bar{x} = f_s(\bar{x})$ satisfies $\bar{x}(0) = 1$.*
4. **false** *if f_s is hereditarily successful and the solution of the equation $\bar{x} = f_s(\bar{x})$ satisfies $\bar{x}(0) = 0$.*

We will call the classifications the semantic status *of s.*

Let us continue the examples in Example 2.13 to see how this definition works.

Example 2.27. *1. The Liar: let $s \in S$ be such that $d(s) = \neg s$. Recall that $f_s : \{0,1\} \to \{0,1\}$ and $f_s(x) = 1 - x$. This function does not have a fixed point, since $f_s(0) = 1$ and $f_s(1) = 0$. Therefore, s is paradoxical.*

2. The Truth Teller: let $s \in S$ be such that $d(s) = s$. We have $f_s : \{0,1\} \to \{0,1\}$ and $f_s(x) = x$. 0 and 1 are both fixed points of f_s. Therefore, s is hypodoxical.

3. Intuitively unproblematic case: Let $s_1, s_2 \in S$ and $d(s_1) = s_2$ and $d(s_2) = \top$. We have $f_{s_1} : \{0,1\}^2 \to \{0,1\}^2$ with the following table:

(x, y)	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$f_{s_1}(x, y)$	$(1,1)$	$(0,1)$	$(1,1)$	$(0,1)$

and $f_{s_2} : \{0,1\} \to \{0,1\}$ with $f_{s_2}(x) = 1$. Therefore, both f_{s_1} and f_{s_2} are successful, and hence they are both hereditarily successful. Moreover, the fixed point of f_{s_1} is $(1,1)$, and the fixed point of f_{s_2} is 1. Therefore, both s_1 and s_2 are true.

4. Liar circle: let $s_0, s_1, \ldots, s_n \in S$ be such that $d(s_i) = s_{i+1}$ for $i < n$, and $d(s_n) = \neg s_0$. Then $f_{s_0} : \{0,1\}^{n+1} \to \{0,1\}^{n+1}$ with $f_{s_0}(\bar{x}) = (x_1, x_2, \ldots, x_n, 1 - x_0)$. We prove that this function does not have a fixed point.

Proof. Assume that $\bar{x}$ is a fixed point of f_{s_0}. Then we have $x_i = x_{i+1}$ for all $i < n$ and $x_n = 1 - x_0$. Therefore, we have $x_0 = x_1 = \cdots = x_n$. Thus, we have $x_0 = 1 - x_0$, but this is impossible as $x_0 \in \{0, 1\}$. Therefore, f_{s_0} does not have a fixed point. □

Therefore, s_0 is paradoxical. Similarly, we can show that $s_1, \ldots, s_n$ are all paradoxical.

5. *Truth Teller circle: let $s_0, s_1, \ldots, s_n \in S$ be such that $d(s_i) = s_{i+1}$ for $i < n$, and $d(s_n) = s_0$. Then $f_{s_0} : \{0,1\}^{n+1} \to \{0,1\}^{n+1}$ with $f_{s_0}(\bar{x}) = (x_1, x_2, \ldots, x_n, x_0)$. One can easily see that there are two fixed points of f_{s_0}, namely $(0, 0, \ldots, 0, 0)$ and $(1, 1, \ldots, 1, 1)$. Therefore, s_0 is hypodoxical. Similarly, $s_1, \ldots, s_n$ are all hypodoxical.*

6. *Yablo's paradox: let $\{s_i \mid i \in \omega\} \subseteq S$ be such that $d(s_i) = \bigwedge_{i<j} s_j$ for $i < n$, and $d(s_n) = \neg s_0$. Then $f_s : \{0,1\}^\omega \to \{0,1\}^\omega$ with the j-th coordinate of f_{s_0} given by $f_{s_0}(\bar{x})(j) = min\{1 - x_i \mid j \leq i\}$.*

   We show that this function does not have a fixed point.

   Proof. Assume that $\bar{x}$ is a fixed point of f_{s_0}. Then $x_j = f_{s_0}(\bar{x})(j) = min\{1 - x_i \mid j \leq i\}$ for all $j \in \omega$.

   If $x_0 = 1$, then, since $x_0 = min\{1 - x_i \mid 0 \leq i\}$, we have $x_i = 0$ for all $i > 0$. Then $0 = x_1 = min\{1 - x_i \mid 1 \leq i\} = 1$, contradiction.

   If $x_0 = 0$. Then there exists j such that $x_j = 1$. Then we have $x_i = 0$ for all $i > j$. However, then $0 = x_{j+1} = min\{1 - x_i \mid j + 1 \leq i\} = 1$, contradiction again.

   Therefore, f_{s_0} does not have a fixed point. □

   Therefore, s_0 is paradoxical. Similarly, we can show that $s_1, \ldots, s_n$ are all paradoxical.

This theory aligns very well with how we classify a sentence in practice. Take Yablo's paradox as an example: we first decide which sentences are relevant here (formally, we determine the variable set), and then we assume that there is a way to assign a truth value for all the relevant sentences in the paradox (we assume f_{s_0} has a fixed point), and we show that a contradiction can be derived whatever value we assign to the first sentence (we prove that f_{s_0} has no fixed point by contradiction). Therefore, the way we derive the contradiction in natural language exactly parallels the way we derive the contradiction in the above proof.

One might note that Definition 2.26 does not cover all the possibilities of f_s. Namely, we did not give a name for the case where f_s is successful, but not hereditarily successful. What can we say about them? In fact, we will show in Section 5.1 that these are sentences which depend on a hypodoxical sentence, and the existence of such sentences would give us a sense in which hypodoxical sentences are also paradoxical. However, let us continue with the study of truth first.

2.3 Properties of Truth

In this section, we first explicate and dispel the concern that the definition of truth might not be coherent. Then, we study some properties of true sentences according to our classification.

Let T be the set of all true sentences in a sentence system (S, d). Is T coherent; or, in other words, could there be true sentences which disagree with each other? There are two situations that we would say intuitively that T is not coherent. Firstly, there might be $s_1, s_2 \in T$, where s_2 depend on s_1, but the ascription function of s_2 says s_1 is false. Secondly, there might be $s_1, s_2 \in T$, such that they do not depend on each other, but there exists s_3 such that s_1 and s_2 depend on s_3. In this case, the ascription functions of s_1 and s_2 would each give a truth value for s_3, but it seems like these values might be different.

However, we can show that these two situations can never happen by the following lemma.

Lemma 2.28. *Let $s_1, s_2 \in T$, and $s_3 \in V(s_1) \cap V(s_2)$. Then the naive truth value of s_3 as determined by f_{s_1} is the same as the naive truth value of s_3 as determined by f_{s_2}.*

Remark 2.29. *Note that the first situation is a special case of this lemma where $s_3 = s_1$.*

Proof. Since $s_1, s_2 \in T$, we have f_{s_1} and f_{s_2} are hereditarily successful. Since $s_3 \in V(s_1) \cap V(s_2)$, we have that f_{s_3} agree with f_{s_1} and f_{s_2}. In particular, the naive truth value of s_3 as determined by f_{s_3} is the same as the naive truth value of s_3 as determined by f_{s_1} or f_{s_2}. Thus, the naive truth value of s_3 as determined by f_{s_1} is the same as the naive truth value of s_3 as determined by f_{s_2}. □

Without the coherency concern, let us study more about the properties of true sentences. An interesting question to ask about this theory is its behaviour with respect to classical tautologies. Are all classical tautologies in T? This is not the case. For instance, consider L_1, which says "L_1 is not true", while L_2 says "L_1 is true or L_1 is not true". This example can be modeled in the formal language by $s_1, s_2 \in S$ where $d(s_1) = \neg s_1$ and $d(s_2) = s_1 \vee \neg s_1$. Then one can check that f_{s_2} does not have any fixed point, so s_2 is paradoxical, although it is a classical tautology. In fact, we have:

Lemma 2.30. *If s is true or false, then for all $t \in V_s$, t is also true or false. Moreover, the truth assignment for t agrees with the truth value it obtains from f_s in the following sense: if the naive truth value of $t \in V_s$ is 0 according to f_s, then it is false; and if the naive truth value of $t \in V_s$ is 1 according to f_s, then it is true.*

Proof. This is a straightforward consequence of Lemma 2.20. □

Note that the other direction of this lemma holds as well — if s only depends on true or false sentences, then s itself is also either true or false:

Lemma 2.31. *Let $s \in S$ be a sentence in the sentence system (S, d). If all $t \in D_s$ are either true or false, then s is either true or false.*

Proof. This follows from Lemma 2.22. □

Thus, we see that classical tautologies that depend on problematic sentences are not classified as true or false. What about classical tautologies that depend only on true or false sentences? We have:

Lemma 2.32. *Let $s \in S$ be a sentence in the sentence system (S, d), and all sentences in D_s are either true or false. If $d(s)$ is a classical tautology, then s is true; and if $d(s)$ is a classical contradiction, then s is false.*

Proof. If $d(s)$ is a classical tautology, $[\![d(s)]\!]_v = 1$ for all valuations v. Moreover, since all sentences in D_s are either true or false, we have f_s is hereditarily successful by lemma 2.22, and the fixed point of f_s satisfies $x(0) = 1$. Therefore, s is true. Similar for the case where $d(s)$ is a classical contradiction. □

However, true sentences are more than just classical tautologies (and false sentences are more than just classical contradictions):

Example 2.33. *Let $s_1, s_2 \in S$ where $d(s_1) = (\neg s_1 \wedge s_2) \vee (s_1 \wedge \neg s_2)$ and $d(s_2) = (s_1 \wedge s_2) \vee (\neg s_1 \wedge \neg s_2)$. We have $V_{s_1} = \{s_1, s_2\}$ and $V_{s_2} = \{s_2, s_1\}$.*

Then one can see that f_{s_1} is the function with the following table:

(x, y)	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$f_{s_1}(x, y)$	$(0,1)$	$(1,0)$	$(1,0)$	$(0,1)$

and f_{s_2} is the function with the following table:

(x, y)	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$f_{s_2}(x, y)$	$(1,0)$	$(0,1)$	$(0,1)$	$(1,0)$

Therefore, $(1, 0)$ is the only fixed point of f_{s_1}, and $(0, 1)$ is the only fixed point of f_{s_2}. Both of them give s_1 the naive truth value of 1 and s_2 the naive truth value of 0. Therefore, s_1 is true and s_2 is false.

However, s_1 is not a classical tautology, and s_2 is not a classical contradiction.

These kinds of sentences correspond to the logic puzzles: let there be a village where all the inhabitants are either knaves (who always lie) or knights (who also always tell the truth). We meet two people in the village, and one of them says "I am a knave and the other person is a knight, or the other way around", while the other says "either both of us are knaves, or both of us are knights". Do we have enough information to determine who they are? I believe that the answer is yes — the first person is a knight and the second is a knave — and the reasoning is exactly the same as the reasoning we used in the above example. Also, recall that this is the example I gave in Section 1.2.

2.4 Sentence Systems and Isomorphism

In this section, I discuss how well the theory we have developed behave with respect to the sentence system. Specifically, we will prove that for any sentence $\phi \in Sent(\mathcal{L}_S)$, we can always assume that there is an s in a sentence system (S, d) such that $d(s) = \phi$. Moreover, the semantic status of a sentence is independent of the sentence system it resides in.

Firstly, let us begin with a seemingly innocuous issue. When I discuss, for example, the Liar paradox, I take an s in a sentence system (S, d) such that $d(s) = \neg s$. However, do we just assume that such a letter s exists in all sentence systems? This question might sound trivial — it seems like one can just use a sentence system that contains the Liar paradox when one is interested in the Liar paradox, and use a potentially different sentence system that contains other sentences when one is interested in other sentences. However, this is not as straightforward as it seems, especially after we have developed a theory of semantic status classification: what if a different choice of S yields a different semantic status for a sentence? For example, could it be that a sentence is judged paradoxical in one system but true in another[3]? We need to formally show that this can never happen: the same sentence will have the same semantic status in every sentence system it resides in.

Note that two different sentence systems might use different propositional letters to represent the same sentence in the natural language, so that they are not literally the same. For example, let $S_1 = \{s_1\}$ and $d_1(s_1) = \neg s_1$, and $S_2 = \{s_2\}$ and $d_2(s_2) = \neg s_2$, where $s_1 \neq s_2$. Then s_1 and s_2 are not the same syntactical object, but clearly they are both representations of the Liar paradox. In the end, we should prove that these sentences receive the same semantic status in the two systems. Therefore, we need to define what it means for two sentences from different sentence systems to be the same.

Intuitively, this should mean that they depend on the same sentences and are built from those sentences in the same way. Note that this definition looks circular as it still uses the notion of sameness. However, one can already give a formal definition of sameness using this idea by requiring the denotation pattern for all sentences that they depend on to be the same.

We first define the substitution of propositional letters in a sentence.

Definition 2.34 (Substitution)**.** *$\phi(x_0, x_1, \ldots, x_\alpha)$ and $\psi(y_0, y_1, \ldots, y_\alpha)$ be two sentences in sentence systems (S_1, d_1) and (S_2, d_2), respectively[4]. The variables x_i's and y_j's are propositional letters that ϕ and ψ contain, respectively. Let $\sigma : \{x_0, x_1, \ldots, x_\alpha\} \to \{y_0, y_1, \ldots, y_\alpha\}$ be a bijection. Let $\bar{x} = (x_0, x_1, \ldots, x_\alpha)$ and $\bar{y} = (y_0, y_1, \ldots, y_\beta)$.*

We write $\psi(\bar{y}) = \sigma(\phi(\bar{x}))$ if $\psi(\bar{y})$ can be obtained from $\phi(\bar{x})$ by substituting x_j for $\sigma(x_j)$ for all $j \in \{0, 1, \ldots, \alpha\}$. In this case, we say that ψ is a substitution instance *of ϕ with*

[3]This question is especially involved when we discuss in Section 5.3 the classical variants of the theory I proposed.

[4]Note that α is an ordinal that might not be finite.

respect to σ.

For an easy example, let $\phi(x_0) = x_0$ and $\psi(y_0) = y_0$. Then by the only $\sigma : \{x_0\} \to \{y_0\}$ we can see that $\psi(y_0) = \sigma(\phi(x_0))$. That is, if we substitute x_0 for y_0, then ψ becomes ϕ.

We then define when two sentence systems are the same:

Definition 2.35 (Isomorphism of Sentence Systems)**.** *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. We say that they are* isomorphic *if there exists a bijection $\sigma : S_1 \to S_2$ such that for all $s \in S_1$, $d_2(\sigma(s)) = \sigma(d_1(s))$. We call such σ an* isomorphism *between (S_1, d_1) and (S_2, d_2).*

In words, two sentence systems are isomorphic if they have the same denotation structure, up to renaming the propositional letters by a bijection σ. Clearly, this isomorphism is an equivalence relation on the class of all sentence systems.

For example, consider the sentence systems where $S_1 = \{s_0, s_1, s_2\}$ and $S_2 = \{s'_0, s'_1, s'_2\}$, where $d_1(s_0) = s_1 \wedge s_2$, $d_1(s_1) = \neg s_1$, and $d_1(s_2) = s_1$; and $d_2(s'_0) = s'_1 \wedge s'_2$, $d_2(s'_1) = \neg s'_1$, and $d_2(s'_2) = s'_1$. Then we can see that $\sigma : S_1 \to S_2$ sending s_i to s'_i is a bijection such that $d_2(\sigma(s)) = \sigma(d_1(s))$. Indeed, there is really no distinction between these two sentence systems, excepts that s_i in S_1 is called s'_i in S_2.

Note that this is a stricter condition than simply requiring their dependence patterns to be the same. For example, $S_3 = \{t_0, t_1, t_2\}$ where $d_3(t_0) = t_1 \vee t_2$, $d_3(t_1) = t_1$, and $d_3(t_2) = \neg t_1$ is not isomorphic to S_1 or S_2, although they have the same dependence pattern: for example, in S_1 we have that s_0 depends on to s_1 and s_2, while s_2 and s_1 depends on s_1; and the same pattern holds for S_3, that t_0 depends on to t_1 and t_2, while t_2 and t_1 depends on t_1. This is because, for example, $d_3(t_0)$ is the disjunction of the sentences it depends on, while $d_1(s_0)$ is their conjunction, and clearly they should not be seen as representing the same sentence.

We can also say when a system is inside another system:

Definition 2.36 (Embedding)**.** *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. We say that S_1 is* embeddable *into S_2 if there exists a injection $\sigma : S_1 \to S_2$ such that for all $s \in S_1$, $d_2(\sigma(s)) = \sigma(d_1(s))$. We call such σ an* embedding *of (S_1, d_1) into (S_2, d_2).*

Therefore, an isomorphism is just an embedding that is also a bijection. A special case of an embedding is when a system is a subsystem of another system. In this case, we can simply take the identity function as the embedding.

Definition 2.37 (Extension and Subsystem)**.** *Let (S, d) be a sentence system. Then we call (S', d') an* extension *of (S, d) if $S \subseteq S'$ and $d'(t) = d(t)$ for all $t \in S$. In this case, we call (S, d) a* subsystem *of (S', d').*

An equivalent definition for embedding is to say that (S_1, d_1) is embeddable into (S_2, d_2) if it is isomorphic to a subsystem of (S_2, d_2):

Lemma 2.38. *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. Then (S_1, d_1) is embeddable into (S_2, d_2) if and only if (S_1, d_1) is isomorphic to a subsystem of (S_2, d_2).*

Proof. $\Rightarrow$: Let $\sigma : S_1 \to S_2$ be an embedding of (S_1, d_1) into (S_2, d_2). Then we can take S_2' as the image of S_1 under σ, and d_2' as the restriction of d_2 to S_2'. Then it is easy to see that (S_2', d_2') is a subsystem of (S_2, d_2) and (S_1, d_1) is isomorphic to (S_2', d_2').

$\Leftarrow$: Let (S_2', d_2') be a substructure of (S_2, d_2) such that (S_1, d_1) is isomorphic to (S_2', d_2'). Then we can define $\sigma : S_1 \to S_2$ by $\sigma(s) = s'$ where $s' \in S'$ is the image of s under the isomorphism. Then it is easy to see that σ is an embedding of (S_1, d_1) into (S_2, d_2). □

An important example of a subsystem is the following. Let s be a propositional letter in the sentence system (S, d). Moreover — recall that d is a function from S to $Sent(\mathcal{L}_S)$ — let us denote the set $\{\phi \in Sent(\mathcal{L}_S) \mid \phi = d(t) \text{ for some } t \in V_s\}$ as $d(V_s)$. Then, by definition of the V_s as the transitive closure of the propositional letters generated by s and the dependence relation, we can easily see that $d(V_s) \subseteq V_s$. Therefore, $(V_s, d|_{V_s})$ is also a sentence system, where $d|_{V_s}$ is the restriction of d to V_s. This is clearly a subsystem of (S, d).

Now we can define when two sentences are the same. Though we still call them isomorphic instead of "same", because the latter might lead to the confusion of viewing them as the same syntactical objects.

Definition 2.39 (Isomphism of Sentences)**.** *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. Let $s_1 \in S_1$ and $s_2 \in S_2$. We say that s_1 and s_2 are* isomorphic *if $(V_{s_1}, d_1|_{V_{s_1}})$ and $(V_{s_2}, d_2|_{V_{s_2}})$ are isomorphic under an isomorphism $\sigma : V_{s_1} \to V_{s_2}$ such that $\sigma(s_1) = s_2$.*

Note that in this case, we can simply identify $t \in V_{s_1}$ with $\sigma(t) \in V_{s_2}$:

Lemma 2.40. *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. Let $s_1 \in S_1$ and $s_2 \in S_2$ be two sentences that are isomorphic. Then there exists sentence systems S_1' and S_2' such that S_1' is isomorphic to S_1 and S_2' is isomorphic to S_2, and there exists $t \in S_1' \cap S_2'$ such that t is isomorphic to s_1 and s_2.*

Proof. Simply rename s_1 and s_2 to t in S_1 and S_2, respectively (where $t \notin S_1 \cup S_2$, so it is an unused propositional variable). □

Isomorphic sentences have the same ascription function:

Lemma 2.41. *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. Let $s_1 \in S_1$ and $s_2 \in S_2$ be isomorphic sentences. Then $f_{s_1} = f_{s_2}$.*

Proof. The ascription function of a sentence is clearly invariant under renaming the variables occurring in the sentence. □

Therefore, we see that the semantic status of a sentence determined by this theory is invariant under the choice of sentence systems.

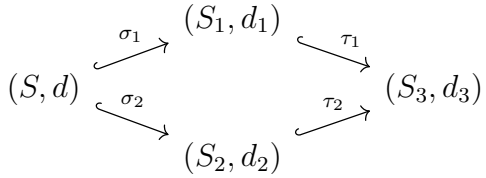
Corollary 2.42. *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. Let $s_1 \in S_1$ and $s_2 \in S_2$ be isomorphic sentences. Then s_1 and s_2 have the same semantic status: s_1 is true (false, paradoxical, hypodoxical) iff s_2 is true (false, paradoxical, hypodoxical).*

Proof. This is clear as these semantic statuses only depend on the ascription function. □

A related question is about the classification. According to Definition 2.26, we only classified the semantic status of propositional letters in a sentence system (S, d). Thus, if one is interested in the semantic status of $s_0 \wedge s_1$, where $s_0, s_1 \in S$, one needs to first find an $s \in S$ such that $d(s) = s_0 \wedge s_1$, and then classify s. Are there always such an s in a sentence system (S, d)? In a finite propositional language, one can simply assume that S is countable and $d : S \to Sent(\mathcal{L}_S)$ is a bijection, because in such languages the cardinality of the set of propositional letters is the same as the cardinality of the set of sentences. However, in an infinite propositional language, this is not the case: the set of sentences has a strictly greater cardinality. Therefore, in any sentence system (S, d), there is always a sentence $\phi \in Sent(\mathcal{L}_S)$ such that there is no $s \in S$ with $d(s) = \phi$. However, because of Lemma 2.42, we can always just work in a larger system which has a name for every sentence in the original system, and this process can be repeated indefinitely.

On top of that, if (S_1, d_1) and (S_2, d_2) have some isomorphic sentences, then we can always find a system (S_3, d_3) such that (S_1, d_1) and (S_2, d_2) are embeddable into (S_3, d_3), and the isomorphic sentences in S_1 and S_2 become the same sentences in S_3. This means that even though we might use different sentence systems while studying different sentences, we can in fact assume that they are all in the same sentence system.

Lemma 2.43. *Let (S_1, d_1) and (S_2, d_2) be two sentence systems. Let (S, d) be a sentence system embeddable into (S_1, d_1) and (S_2, d_2), with embeddings $\sigma_1 : S \to S_1$ and $\sigma_2 : S \to S_2$. Then there exists a sentence system (S_3, d_3) such that (S_1, d_1) and (S_2, d_2) are embeddable into (S_3, d_3), via $\tau_1 : S_1 \to S_3$ and $\tau_2 : S_2 \to S_3$, such that $\tau_1 \circ \sigma_1 = \tau_2 \circ \sigma_2$.*



Proof. Let $S_3 = S \sqcup (S_1 \setminus \sigma_1(S)) \sqcup (S_2 \setminus \sigma_2(S))$. Define d_3 on S_3 as follows: for all $s \in S$, $d_3(s) = d(s)$; for all $s \in S_1$, $d_3(s) = d_1(s)$; and for all $s \in S_2$, $d_3(s) = d_2(s)$. Define $\tau_1 : S_1 \to S_3$ as the inclusion map on $S_1 \setminus \sigma_1(S)$ and $\tau_1(\sigma(s)) = s$ for all $s \in S$, and $\tau_2 : S_2 \to S_3$ as the inclusion map on $S_2 \setminus \sigma_2(S)$ and $\tau_2(\sigma(s)) = s$ for all $s \in S$. It is easy to see that τ_1 and τ_2 are embeddings of (S_1, d_1) and (S_2, d_2) into (S_3, d_3) and satisfy $\tau_1 \circ \sigma_1 = \tau_2 \circ \sigma_2$. □

In words, we can always paste (S_1, d_1) and (S_2, d_2) together along the common subsystem (S, d)[5]:

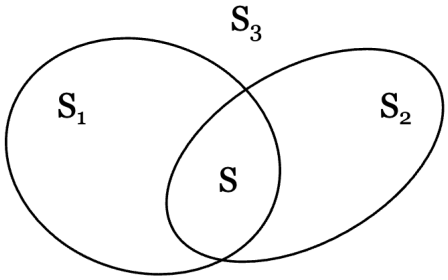


Figure 2.1: Pasting S_1 and S_2 along S.

Collecting all the results above, in practice, for any sentence system and any sentences that we are interested in, we can always assume that there are propositional letters in the sentence system that represent them, without worrying about whether it would influence the semantic status of any sentence in the language. Though strictly speaking, we sometimes need to move to a larger system.

This is a very desirable feature, since the sentences that s does not depend on either directly or indirectly should be irrelevant to the semantic status of s. Moreover, as a theory of truth that aims to guide our classification of sentences in the natural language, it should be able to assume that all sentences of interests are in the language system.

I would later argue that a naive modification to the Revision Theory of truth fails to respect this feature, because the revision rule is globally applied to all sentences in the sentence system, making the classification sensitive to which sentences are included in the system.

2.5 Interpretation of the Language*

We now discuss the limitation of this language and how to interpret a theory formed in this formal language.

Note that this section and the next is a supplementary discussion to the rest of the thesis, and the readers may safely skip both sections on first reading without affecting the understanding of the rest of the thesis. Nevertheless, these two sections are important for understanding the scope of the theory and how to use it to guide the classification of sentences in natural language in a rigorous way.

Firstly, as a propositional language, its expressive power is limited — it cannot look into the detailed structure of a sentence in the sense that it does not have function and

[5]In fact, this is the amalgamation property. See, for example, (Tent, Ziegler, 2012) for a more detailed discussion of this property.

predicate symbols. Therefore, it is not possible for us to distinguish between, for example, "$1 + 1 = 2$", "$1 + 2 = 3$", and "grass is green". In a first order language, these can be represented formally as "$1 + 1 = 2$", "$1 + 2 = 3$", and "$Green(grass)$", and we can give a model where these are indeed true. In this propositional language, however, we cannot distinguish between these truths that depend only on worldly facts. All of them will simply be represented as $\top$. Similarly, all falsehoods that depend only on worldly facts will be represented as $\bot$.

This is a serious limitation for those who want to use the language to study, for example, arithmetical truth, because only after knowing the truth of a sentence can we then represent the sentence in our formal language. However, this is not a problem for us, because what we are interested in here using this language is only the truth and falsity of the sentences which depends also on other sentences.

Thus, the choice of the language is the reflection of a simplified purpose: assume that we have already known the worldly facts; how do we determine the truth value of sentences that refer to the expressions of these facts, or those that refer to each other? In a word, it is not of our current interest to use the language to study worldly facts.

Nevertheless, it would indeed be an interesting project to develop these ideas in a first order language, for example, in the language of arithmetic, where we could use the theory to both determined arithmetical truth and the truth of sentences that refer to each other, thus giving a theory of truth in a first order language. However, as observed by (Beringer and Schindler, 2017) this is much harder than doing it in this propositional language. We will discuss this in more detail in Chapter 6.

Aside from the limitations resulting from the choice of a propositional language, there are also some potential difficulties resulting from the use of the denotation function. We should be able to apply a theory of truth formed in this language to the natural language, otherwise it would be unclear why one should care about it in such a limited propositional language. For example, suppose I classify the sentence s_0 with $d(s_0) = \neg s_0$ as paradoxical. It is not of much interest to just claim that what we do is classifying a sentence s_0 in the infinite propositional language as paradoxical. Unlike the first order language — which is of great significance because of its fruitful application in mathematics — this propositional language is just designed to study the paradox. Therefore, we want to interpret this formal result as saying that the Liar sentence "this sentence is not true" is paradoxical. However, deviating from the first order language again, this language lacks a truth predicate, so it is not as clear how to apply a theory of truth designed in this language to answer questions about truth in natural language.

In the following, I will give examples about truth that one might want to ask and discuss the challenges one might encounter while trying to answer them using a theory in this language. Then I propose two suggestions on how to deal with these challenges.

Firstly, one might wonder whether this language can be used to represent iterated truth.

For example, how do we distinguish between "grass is green", "$True$(grass is green)", and "$True(True$(grass is green))"? In fact, this is quite straightforward. We express[6] "grass is green" by $\phi_0 = \top$, "$True$(grass is green)" by s_1 such that $d(s_1) = \phi_0 = \top$, and "$True(True$(grass is green))" by s_2 such that $d(s_2) = s_1$. Therefore, the denotation function d is really understood as taking the place of truth predicate in front of every propositional letter: suppose $d(s) = \phi(s_0, \ldots, s_\alpha)$, where s_i is interpreted as a sentence L_i, then s represents the sentence $\phi(True(L_0), \ldots, True(L_\alpha))$. Note that this is consistent with our previous examples: $d(s_0) = \neg s_0$ is interpreted as the Liar sentence L_0 saying "L_0 is not true", and $d(s_1) = s_1$ is interpreted as the Truth Teller sentence L_1 saying "L_1 is true".

There is an immediate problem with this interpretation: what about sentences that contain connectives inside the scope of a truth predicate? For example, "the negation of this sentence is true" or "it is true that L_1 and L_2" where L_1 and L_2 are just some normal sentences like "grass is green". These sentences can be faithfully translated into first order language as "$True(\ulcorner \neg L \urcorner)$" and "$True(\ulcorner L_1 \wedge L_2 \urcorner)$", respectively. Given our interpretation, there is no way to represent them in the propositional language, since the scope of a truth predicate only ever contains the name of a single sentence.

One might propose that this is not a problem because we can freely interpret our formal results into natural language. For example, by classifying the sentence s_1 such that $d(s_1) = \neg s_1$ as paradoxical, we can conclude that both L_1 saying "L_1 is not true" and L_1' "the negation of L_1' is true" are paradoxical. In general, the answer to (1) whether a sentence like "L_i is not true" is true is the same as the answer to (2) whether a sentence like "the negation of L_i is true"; and the answer to (1) whether a sentence like "it is true that L_i and L_j" is true is the same as the answer to (2) whether a sentence like "L_i is true and L_j is true" is true.

However, this freedom of interpretation comes at the cost of assuming some metalinguistic principles — on what grounds can we claim that the answer to (1) is the same as the answer to (2)? The most straightforward answer is that we assume the truth predicate commutes with logical connectives. Therefore, in the above examples, the sentence "the negation of this sentence is true" is just equivalent to the sentence "this sentence is not true", and the sentence "it is true that L_1 and L_2" is equivalent to the sentence "L_1 is true and L_2 is true". Nevertheless, these assumptions are too strong for the specific theory I present in this thesis. Recall from the previous paragraph that we are able to represent iterated truth in the propositional language. We will see later that the theory of truth I propose will suggests that the truth predicate is idempotent: for any sentence L, we would have "L is true" if and only if "it is true that L is true". Unfortunately, this would be inconsistent with the assumption that the truth predicate commutes with all logical connectives (see Halbach and Leigh, 2024, Chapter 11; Halbach, 2014, Chapter 13).

[6]Note that $\top$ is not a propositional variable.

There are two ways out of this dilemma — either by committing to weaker metalinguistic principles or admitting the expressive limits of our language. First, we do not really need to assume that the truth predicate commutes with all logical connectives. Instead, we can just assume that (i) the truth predicate is idempotent, and (ii) the truth predicate commutes with logical connectives *inside the scope of a truth predicate*. For example, instead of assuming that "L is not true" if and only if "the negation of L is true", we can assume that "it is true that L is not true" if and only if "it is true that the negation of L is true". This can still solve our problem because it would indeed allow us to conclude that the answer to (1) whether a sentence like "the negation of this sentence is true" is true is the same as the answer to (2) whether a sentence like "this sentence is not true" is true. Moreover, these assumptions are indeed milder — formulated in the most straightforward way in the first order language, they are implied by the Kripke–Feferman theory of truth (see Halbach and Leigh, 2022; Halbach, 2014: Chapter 15).

However, one might not want to commit to any such metalinguistic principles. For example, it is not at all clear that the principle that the truth predicate commutes with logical connectives should be given up in favour of the principle that the truth predicate is idempotent (see Halbach and Leigh, 2022; Halbach, 2014, Chapter 14). In this case, I suggest that one should just admit the expressive limits of the language. One cannot expect to have a fruitful or even consistent theory by giving completely free interpretations to the formal results of this theory. In particular, I suggest that one interpret the language in the way presented in the previous paragraph — understand d as taking the place of truth predicate in front of every propositional letter — because in this way we can at least cover the most important examples of semantic paradoxes in the literature.

Admittedly, both are not ideal solutions, but the reason we have these limitations is that we are working only in a propositional language, and we do not have the expressive power to express the truth predicate. However, as we have seen, the theory of truth built from it can already be very fruitful in guiding us to understand most of the significant semantic paradoxes, such as the Liar paradox, the Truth Teller paradox, and the Yablo's paradox. Moreover, as I will discuss in the Chapter 6, the ideas have the potential to be developed in a first order theory, where no metatheoretical assumptions are needed.

2.6 Interpretation of the Theory*

We now move on to discuss how to interpret the theory we have developed in this propositional language. As discussed in the previous section, we have to either assume some metatheoretical assumptions or admit that the expressive power of the language has some limitations. We now show that if one is willing to accept those assumptions, then the theory can also prove that the formal counterparts of these assumptions hold.

Firstly, we need another metatheoretical assumption that the truth predicate is

idempotent. We just showed in Section 2.4 that if we are interested in whether a sentence ϕ is true, we can always find a sentence s in the sentence system (S, d) such that $d(s) = \phi$. However, recall from Section 2.5 that this s is interpreted as "ϕ is true". Thus, if we later decide that s is true, strictly speaking, we are not saying that ϕ is true, but that "ϕ is true" is true. Hence, if we want to conclude from the formal theory that the sentence ϕ itself is true, we need another assumption that the truth predicate is idempotent: "ϕ is true" if and only if "it is true that ϕ is true". As we have already discussed, this assumption is a consequence of the Kripke-Feferman theory of truth. In addition, there is no inconsistency with our other metatheoretical assumptions.

Hence, we have made the following assumptions in the metatheory: the truth predicate is idempotent and it commutes with the logical connectives inside the scope of a truth predicate.

Now we want to show that the theory is really coherent with these assumptions in the sense that these assumptions can be expressed and proved within the theory. Note a potential confusion: what is the point of proving, for example, that the truth predicate is idempotent in a theory that already assumes it? Is it not circular and so meaningless?

Firstly, it is not strictly speaking circular because the assumption is made in the metatheory, not within the theory. Moreover, the theory can be seen as a model of the metatheory. It is rather desirable that the model can express and prove some statements that are true in the metatheory. For example, first order logic is a model of classical mathematical reasoning. Within first order logic, we express and prove that "proof by contradiction" holds, while we feel free to prove this fact using the method of proof by contradiction in the metalanguage. As another example, in set theory, we prove that the "induction principle" holds by using mathematical induction. There is really no circularity here — instead, we are proving that the theory built in the formal language is indeed a very good model of the metatheory.

Let us first show that the truth predicate commutes with negation inside the scope of truth. Recall that this means "it is true that L is not true" if and only if "it is true that the negation of L is true".

Lemma 2.44 (Commuting with Negation)**.** *Let $\phi \in Sent(\mathcal{L}_S)$. Let $s_1, s_2 \in S$ be such that $d(s_1) = \phi$ and $d(s_2) = \neg s_1$. Let $s_3 \in S$ be such that $d(s_3) = \neg\phi$. Then the semantic status of s_2 is the same as that of s_3.*

Remark 2.45. *Under other metatheoretical assumptions, s_1 represents "ϕ is true", s_2 represents "ϕ is not true", and s_3 represents "the negation of ϕ is true". Thus, we will have "it is true that ϕ is not true" if and only if "it is true that the negation of ϕ is true".*

Proof. We first assume that s_2 has a certain status and prove that s_3 has the same.

1. s_2 is paradoxical. Assume for a contradiction that s_3 has a fixed point. Then this fixed point induces a valuation on the propositional variables occurring in ϕ. This

fixed point clearly induces a fixed point for s_1, which in turn can easily be extended to a fixed point for s_2. This is a contradiction.

2. s_2 is hypodoxical. Let $\bar{x}$ and $\bar{y}$ be two distinct fixed points of f_{s_2}. Let v_x and v_y be the valuation induced by $\bar{x}$ and $\bar{y}$ respectively. I claim that there must be some $t \in V(s_1)$ such that $v_x(t) \neq v_y(t)$. Otherwise, by 2.4 we must have $v_x(s_2) = [\![d(s_2)]\!]_{v_x} = [\![\neg s_1]\!]_{v_x} = [\![\neg s_1]\!]_{v_y} = v_y(s_2)$. This means that v_x and v_y are the same valuation on $V(s_2)$, which could happen only if $\bar{x} = \bar{y}$. This is a contradiction. Therefore, v_x and v_y induce distinct fixed points for f_{s_1}. Let $\bar{w}$ and $\bar{v}$ be two distinct fixed points of f_{s_1}. Then changing v_0 and w_0 to $1 - v_0$ and $1 - w_0$ gives two distinct fixed points for f_{s_3}. Thus, s_3 is also hypodoxical.

3. s_2 is true. Using ideas similar to those in the previous cases, we can see s_3 is also true.

4. s_2 is false. Using ideas similar to those in the previous cases, we can see s_3 is also false.

Next, we need to assume s_3 has a certain status and prove that s_2 has the same. The proof is similar to the above case, so we omit it. □

We now express the other metatheoretical assumptions. However, the proofs are very similar to the above case, so we omit them.

Lemma 2.46 (Idempotence)**.** *Let $\phi \in Sent(\mathcal{L}_S)$. Let $s_1, s_2 \in S$ be such that $d(s_1) = \phi$ and $d(s_2) = s_1$. Then the semantic status of s_1 is the same as that of s_2.*

Lemma 2.47 (Commuting with Conjunction)**.** *Let $\phi, \psi \in Sent(\mathcal{L}_S)$, and $s_1, s_2 \in S$ be such that $d(s_1) = \phi$ and $d(s_2) = \psi$. Let $s_3 \in S$ be such that $d(s_3) = s_1 \wedge s_2$. Then s_3 is true if and only if s_1 and s_2 are both true.*

Lemma 2.48 (Commuting with Disjunction)**.** *Let $\phi_0, \ldots, \phi_\alpha \in Sent(\mathcal{L}_S)$, and $s_1, \ldots, s_\alpha \in S$ be such that $d(s_i) = \phi_i$ for all $i \leq \alpha$. Let $s \in S$ be such that $d(s) = \bigwedge_{i \leq \alpha} s_i$. Then s is true if and only if all s_i's are true.*

Therefore, the theory is indeed coherent with the metatheoretical assumptions we made.

Chapter 3

Comparison with Kripke's Theory of Truth

There are two dominant theories which are also hierarchy-free — Kripke's minimal fixed point theory of truth and the Revision Theory of Truth. In the following two chapters, I compare the proposed theory with these two theories.

Note that Kripke's theory and the Revision Theory are both formulated in the first order language, while the theory I proposed is formulated in an infinitary propositional language. Therefore, I will develop their theories in the infinitary propositional language we use here. Then we can formally compare the extension of truth in the two theories with the extension of truth in our theory. Nevertheless, to illustrate that the difference is not a consequence of the way I formulated the theory in this language, I will argue that the corresponding differences also show up in the first order language.

3.1 Kripke's Theory of Truth in the First Order Language

Let us first review Kripke's original theory in the first order language of arithmetic $\mathcal{L} = \{0, +, \times, s\}$, together with its standard model $\mathbb{N}$. The goal is to extend the language with a truth predicate $True$. We call the extended language $\mathcal{L}^T$ (cf. Kripke, 1975; Field, 2008: Ch3).

The key idea of Kripke's theory is that we determine the extension of the truth predicate step by step: imagine that "we are explaining the word 'true' to someone who does not yet understand it" (Kripke 1975, p. 701) by telling them the T-schema, namely, that we can say a sentence is true if we can assert that sentence itself. At the first step, they would understand whether a sentence that does not involve the notion of truth is true or false. For example, they would grasp that "snow is white" is true because they know they can assert "snow is white". At the second step, they now also know that they

can assert "'snow is white' is true" and hence T-schema applies again to conclude that "'snow is white' is true" is true. "In this manner, the subject will eventually be able to attribute truth to more and more statements involving the notion of truth itself." (p. 701).

Hence, in general, at the beginning, the truth of every sentence is indeterminate. At each step, based on what have been found to be true or false, we use an update rule to decide which previously indeterminate sentences can now be confirmed as true or false. This update procedure goes to the transfinite and eventually reaches a fixed point where no more sentences can be confirmed as true or false.

Formally, we assign to $True$ both an extension S_1, containing (codes of) sentences which are true, and an anti-extension S_2, containing sentences which are false. Note that $S_1, S_2 \subseteq \mathbb{N}$, but as they are just codes of sentences in our language, we can just see them as sets of sentences in $\mathcal{L}^T$. We require $S_1 \cap S_2 = \emptyset$ and the other sentences which are in neither of them are understood as undecided. Given a pair (S_1, S_2), we can evaluate the truth of a sentence $\phi \in Sent(\mathcal{L}^T)$ according to some three-valued logic. This process, taking a pair (S_1, S_2), and then returning a valuation for a sentence ϕ from the values $\{0, \frac{1}{2}, 1\}$, is called a valuation scheme. We present the valuation schemes based on strong Kleene logic and weak Kleene logic[1].

Definition 3.1 (Strong Kleene Valuation Scheme)**.** *The* strong Kleene valuation scheme $V_{SK} : (\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})) \to (Sent(\mathcal{L}^T) \to \{0, \frac{1}{2}, 1\})$ *is defined as follows: let* $(S_1, S_2) \in \mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$ *be a pair of sets of sentences in* $\mathcal{L}^T$*, where* $S_1 \cap S_2 = \emptyset$*. Then we define* $v = V_{SK}(S_1, S_2) : Sent(\mathcal{L}^T) \to \{0, \frac{1}{2}, 1\}$ *recursively as follows: let* $\phi \in Sent(\mathcal{L}^T)$*,*

1. *if* ϕ *is an atomic sentence: in* $\mathcal{L}$*, then* $v(\phi) = 1$ *if* $\mathbb{N} \models \phi$*,* $v(\phi) = 0$ *if* $\mathbb{N} \not\models \phi$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

2. *if* $\phi = True(t)$ *for some numeral* t*, then* $v(\phi) = 1$ *if* $t \in S_1$*,* $v(\phi) = 0$ *if* $t \in S_2$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

3. *if* $\phi = \neg\psi$*, then* $v(\phi) = 1$ *if* $v(\psi) = 0$*,* $v(\psi) = 0$ *if* $v(\phi) = 1$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

4. *if* $\phi = \psi_1 \wedge \psi_2$*, then* $v(\phi) = 1$ *if* $v(\psi_1) = v(\psi_2) = 1$*,* $v(\phi) = 0$ *if one of* $v(\psi_1)$ *and* $v(\psi_2)$ *is* 0*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

5. *if* $\phi = \psi_1 \vee \psi_2$*, then* $v(\phi) = 1$ *if one of* $v(\psi_1)$ *and* $v(\psi_2)$ *is* 1*,* $v(\phi) = 0$ *if* $v(\psi_1) = v(\psi_2) = 0$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

6. *if* $\phi = \exists x\psi(x)$*, then* $v(\phi) = 1$ *if there exists* $a \in \mathbb{N}$ *such that* $v(\psi(a)) = 1$*,* $v(\phi) = 0$ *if for all* $a \in \mathbb{N}$*,* $v(\psi(a)) = 0$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

[1]For a detailed discussion of the motivations and properties of the Kleene logics, see (Priest, 2008: Chapter 7).

7. *if* $\phi = \forall x\psi(x)$*, then* $v(\phi) = 1$ *if for all* $a \in \mathbb{N}$*,* $v(\psi(a)) = 1$*,* $v(\phi) = 0$ *if there exists* $a \in \mathbb{N}$ *such that* $v(\psi(a)) = 0$*, and* $v(\phi) = \frac{1}{2}$ *otherwise.*

The valuation scheme based on weak Kleene logic is defined similarly, by modifying the clauses for $\wedge$, $\vee$, $\neg$, $\exists$, and $\forall$ according to weak Kleene logic.

Definition 3.2 (Weak Kleene Valuation Scheme)**.** *The* weak Kleene valuation scheme $V_{WK} : (\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})) \to (Sent(\mathcal{L}^T) \to \{0, \frac{1}{2}, 1\})$ *is defined as follows: let* $(S_1, S_2) \in \mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$ *be a pair of sets of sentences in* $\mathcal{L}^+$*, where* $S_1 \cap S_2 = \emptyset$*. Then we define* $v = V_{SK}(S_1, S_2) : Sent(\mathcal{L}^T) \to \{0, \frac{1}{2}, 1\}$ *recursively as follows: let* $\phi \in Sent(\mathcal{L}^T)$*,*

1. *if* ϕ *is an atomic sentence: in* $\mathcal{L}$*, then* $v(\phi) = 1$ *if* $\mathbb{N} \models \phi$*,* $v(\phi) = 0$ *if* $\mathbb{N} \not\models \phi$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

2. *if* $\phi = True(t)$ *for some numeral* t*, then* $v(\phi) = 1$ *if* $t \in S_1$*,* $v(\phi) = 0$ *if* $t \in S_2$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

3. *if* $\phi = \neg\psi$*, then* $v(\phi) = 1$ *if* $v(\psi) = 0$*,* $v(\psi) = 0$ *if* $v(\phi) = 1$*, and* $v(\phi) = \frac{1}{2}$ *otherwise;*

4. *if* $\phi = \psi_1 \wedge \psi_2$*, then* $v(\phi) = 1$ *if* $v(\psi_1) = v(\psi_2) = 1$*,* $v(\phi) = \frac{1}{2}$ *if one of* $v(\psi_1)$ *and* $v(\psi_2)$ *is* $\frac{1}{2}$*, and* $v(\phi) = 0$ *otherwise;*

5. *if* $\phi = \psi_1 \vee \psi_2$*, then* $v(\phi) = 0$ *if* $v(\psi_1) = v(\psi_2) = 0$*,* $v(\phi) = \frac{1}{2}$ *if one of* $v(\psi_1)$ *and* $v(\psi_2)$ *is* $\frac{1}{2}$*, and* $v(\phi) = 1$ *otherwise;*

6. *if* $\phi = \exists x\psi(x)$*, then* $v(\phi) = \frac{1}{2}$ *if there exists* $a \in \mathbb{N}$ *such that* $v(\psi(a)) = \frac{1}{2}$*,* $v(\phi) = 0$ *if for all* $a \in \mathbb{N}$*,* $v(\psi(a)) = 0$*, and* $v(\phi) = 1$ *otherwise;*

7. *if* $\phi = \forall x\psi(x)$*, then* $v(\phi) = 1$ *if for all* $a \in \mathbb{N}$*,* $v(\psi(a)) = 1$*,* $v(\phi) = \frac{1}{2}$ *if there exists* $a \in \mathbb{N}$ *such that* $v(\psi(a)) = \frac{1}{2}$*, and* $v(\phi) = 0$ *otherwise.*

Then given a pair (S_1, S_2) of initial attempts for the extension and anti-extension for $True$, we can apply the valuation scheme to obtain a valuation v for any sentence $\phi \in Sent(\mathcal{L}_S)$. The sentences which are assigned 1 would then become our next attempt for the extension of $True$, and those assigned 0 the next anti-extension. This update process is called the Kripke-jump.

Definition 3.3 (Kripke-jump)**.** *Let* V *be a valuation scheme. The* Kripke-jump $\mathcal{J}_V : \mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N}) \to \mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$ *of* V *is defined as follows: let* $(S_1, S_2) \in \mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$ *be a pair of sets of sentences in* $\mathcal{L}^T$*. Then we define* $\mathcal{J}_V(S_1, S_2) = (S_1', S_2')$ *where* $S_1' = \{\phi \in Sent(\mathcal{L}^T) \mid V(S_1, S_2)(\phi) = 1\}$ *and* $S_2' = \{\phi \in Sent(\mathcal{L}^T) \mid V(S_1, S_2)(\phi) = 0\}$*. We say that* S_1' *has truth value* 1 *relative to* (S_1, S_2) *and* S_2' *has truth value* 0 *relative to* (S_1, S_2)*.*

Starting from the empty sets $(\emptyset, \emptyset)$, if we keep applying the Kripke-jump, we will eventually reach a fixed point (S_1, S_2) such that $\mathcal{J}_V(S_1, S_2) = (S_1, S_2)$, as long as the valuation scheme is nice enough[2]. In the following, we assume V is either V_{SK} or V_{WK}. For simplicity of notation, we write $(X, Y) \subseteq (X', Y')$ when $X \subseteq X'$ and $Y \subseteq Y'$.

Theorem 3.4 (Kripkean Truth)**.** *Let $(X_0, Y_0) = (\emptyset, \emptyset)$. For any ordinal α, let $(X_{\alpha+1}, Y_{\alpha+1}) = \mathcal{J}_V(X_\alpha, Y_\alpha)$. If α is a limit ordinal, let $(X_\alpha, Y_\alpha) = \bigcup_{\beta<\alpha}(X_\beta, Y_\beta)$. Then there exists an ordinal β such that $(X_{\beta+1}, Y_{\beta+1}) = \mathcal{J}_V(X_\beta, Y_\beta)$. We denote X_β as T_K, the Kripkean extension of truth.*

Proof. By going through the recursive definition of V, one can routinely check that if $(X, Y) \subseteq (X', Y')$, then $\mathcal{J}_V(X, Y) \subseteq \mathcal{J}_V(X', Y')$. As $(X_0, Y_0) = (\emptyset, \emptyset)$, we have $(X_0, Y_0) \subseteq (X_1, Y_1)$. Then a straightforward induction argument shows that $(X_\alpha, Y_\alpha) \subseteq (X_{\alpha+1}, Y_{\alpha+1})$ for all ordinals α. Since the number of sentences is countable, there must be a β such that $(X_{\beta+1}, Y_{\beta+1}) = \mathcal{J}_V(X_\beta, Y_\beta)$. □

3.2 Kripke's Theory of Truth in the Infinitary Propositional Language

We now develop a Kripkean theory of truth in the infinitary propositional language. The essential point is to define the Kripke-jump, which, given a pair of extension and anti-extension for truth, outputs a revised pair based on them.

We first define the two valuation schemes, which — given an extension and an anti-extension of truth — would allow us to evaluate the truth value of any sentence in the language. Recall that in the theory I proposed, the extension of truth would be a subset of the propositional letters S in the language. Therefore, we also require the extension and anti-extension to be subsets of S, so that in the end we can formally compare the extension of truth in the two theories. Hence, in the propositional language, a valuation scheme would take a pair of subsets of S, and return a valuation for any sentence in the language, which is a function from $Sent(\mathcal{L}_S)$ to $\{0, \frac{1}{2}, 1\}$.

Definition 3.5 (Strong Kleene Valuation Scheme)**.** *The* strong Kleene valuation scheme $V_{SK} : \mathcal{P}(S) \times \mathcal{P}(S) \to (Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\})$ *is defined as follows: let $(S_1, S_2) \in \mathcal{P}(S) \times \mathcal{P}(S)$ be a pair of sets of sentences in S. Then we define $v = V_{SK}(S_1, S_2) : Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\}$ recursively as follows: let $\phi \in Sent(\mathcal{L}_S)$,*

1. *if $\phi \in \{\top, \bot\}$: then $v(\phi) = 1$ if $\phi = \top$, $v(\phi) = 0$ if $\phi = \bot$;*
2. *if $\phi = s$ for some $s \in S$: then $v(\phi) = 1$ if $s \in S_1$, $v(s) = 0$ if $s \in S_2$, and $v(\phi) = \frac{1}{2}$ otherwise;*

[2]For precise conditions, see (Herzberger, 1982b) and (Beringer and Schindler, 2017).

3. *if* $\phi = \neg\psi$, *then* $v(\phi) = 1$ *if* $v(\phi) = 0$, $v(\phi) = 0$ *if* $v(\phi) = 1$, *and* $v(\phi) = \frac{1}{2}$ *otherwise;*

4. *if* $\phi = \psi_1 \wedge \psi_2$, *then* $v(\phi) = 1$ *if* $v(\psi_1) = v(\psi_2) = 1$, $v(\phi) = 0$ *if one of* $v(\psi_1)$ *and* $v(\psi_2)$ *is* 0, *and* $v(\phi) = \frac{1}{2}$ *otherwise;*

5. *if* $\phi = \bigwedge \Phi$, *then* $v(\phi) = 1$ *if for all* $\psi \in \Phi$, $v(\psi) = 1$, $v(\phi) = 0$ *if one of* $v(\psi)$ *is* 0, *and* $v(\phi) = \frac{1}{2}$ *otherwise.*

Remark 3.6. *Note that* S_1 *is the extension of truth, while* S_2 *is the anti-extension. Moreover, clause* 2 *is where these extensions come into play — this is exactly the same situation as in the first order case.*

Definition 3.7 (Weak Kleene Valuation Scheme)**.** *The* weak Kleene valuation scheme $V_{WK} : \mathcal{P}(S) \times \mathcal{P}(S) \to (Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\})$ *is defined as follows: let* $(S_1, S_2) \in \mathcal{P}(S) \times \mathcal{P}(S)$ *be a pair of sets of sentences in* S. *Then we define* $v = V_{WK}(S_1, S_2) : Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\}$ *recursively as follows: let* $\phi \in Sent(\mathcal{L}_S)$,

1. *if* $\phi \in \{\top, \bot\}$*: then* $v(\phi) = 1$ *if* $d(s) = \top$, $v(\phi) = 0$ *if* $d(s) = \bot$*;*

2. *if* $\phi = s$ *for some* $s \in S$*: then* $v(\phi) = 1$ *if* $s \in S_1$, $v(s) = 0$ *if* $s \in S_2$, *and* $v(\phi) = \frac{1}{2}$ *otherwise;*

3. *if* $\phi = \neg\psi$, *then* $v(\phi) = 1$ *if* $v(\phi) = 0$, $v(\phi) = 0$ *if* $v(\phi) = 1$, *and* $v(\phi) = \frac{1}{2}$ *otherwise;*

4. *if* $\phi = \psi_1 \wedge \psi_2$, *then* $v(\phi) = 1$ *if* $v(\psi_1) = v(\psi_2) = 1$, $v(\phi) = 0$ *if one of* $v(\psi_1)$ *and* $v(\psi_2)$ *is* $\frac{1}{2}$, *and* $v(\phi) = 0$ *otherwise;*

5. *if* $\phi = \bigwedge \Phi$, *then* $v(\phi) = 0$ *if for all* $\psi \in \Phi$, $v(\psi) = 0$, *and otherwise it is the same as the above.*

Note that if all propositional letters of a sentence are either in the extension or in the anti-extension of truth, then the weak Kleene valuation and strong Kleene valuation will agree on the truth value of this sentence. Moreover, this will just be the classical valuation of this sentence:

Lemma 3.8. *Let* $\phi \in Sent(\mathcal{L}_S)$ *and* $(S_1, S_2) \in \mathcal{P}(S) \times \mathcal{P}(S)$. *Let* $v : Sent(\mathcal{L}_S) \to \{0, 1\}$ *be a classical valuation such that* $v(s) = 1$ *for all* $s \in S_1$ *and* $v(s) = 0$ *for all* $s \in S_2$.

If all propositional letters occurring in ϕ *are in* $S_1 \cup S_2$, *then we have* $V_{WK}(S_1, S_2)(\phi) = V_{SK}(S_1, S_2)(\phi) = [\![\phi]\!]_v$.

Proof. Induction on the complexity of ϕ. Notice from the definition of the Kleene schemes that when no formula receives the value $\frac{1}{2}$, the valuation schemes agree with the classical valuation. □

We now proceed to define the Kripke-jump.

Definition 3.9 (Kripke-jump)**.** *Let V be a valuation scheme. The* Kripke-jump $\mathcal{J}_V : \mathcal{P}(S) \times \mathcal{P}(S) \to \mathcal{P}(S) \times \mathcal{P}(S)$ *of V is defined as follows: let $(S_1, S_2) \in \mathcal{P}(S) \times \mathcal{P}(S)$ be a pair of sets of sentences in $\mathcal{L}_S$. Then we define $\mathcal{J}_V(S_1, S_2) = (S_1', S_2')$ where $S_1' = \{s \in S \mid V(S_1, S_2)(d(s)) = 1\}$ and $S_2' = \{s \in S \mid V(S_1, S_2)(d(s)) = 0\}$.*

The fixed point theorem holds as in the first order case, so that we can define the Kripkean extension of truth in the infinitary propositional language:

Theorem 3.10 (Kripkean Truth)**.** *Let $(X_0, Y_0) = (\emptyset, \emptyset)$. For any ordinal α, let $(X_{\alpha+1}, Y_{\alpha+1}) = \mathcal{J}_V(X_\alpha, Y_\alpha)$. If α is a limit ordinal, let $(X_\alpha, Y_\alpha) = \bigcup_{\beta<\alpha}(X_\beta, Y_\beta)$. Then there exists an ordinal β such that $(X_{\beta+1}, Y_{\beta+1}) = \mathcal{J}_V(X_\beta, Y_\beta)$. We denote X_β as T_K, the Kripkean extension of truth.*

We denote the Kripkean extension of truth based on the weak Kleene logic as T_K^W and the Kripkean extension of truth based on the strong Kleene logic as T_K^S.

3.3 Supvervaluation Version of Kripke's Theory of Truth

In this section, we point out an undesirable feature of the Kripkean theories based on the Kleene logics, which can be solved by using supervaluation instead of the Kleene logics. I will follow Field (2008: Ch 10) in the presentation of this theory in the first order language.

An unsatisfactory feature of the Kripkean theories is that they sacrifice too many of the classical tautologies. For example, consider a sentence s_1 such that $d(s_1) = s_1 \vee \neg(s_1)$. Neither Kripkean theories classifies s_1 as true, because it is not in the minimal fixed point — at the initial stage, s_1 receives value $\frac{1}{2}$, and it remains so in all later stages because $v(\frac{1}{2} \vee \frac{1}{2}) = \frac{1}{2}$ in both valuation schemes. Note that the original Kripkean theories in the first order language have this problem as well — a sentence $L_1 = True(\ulcorner L_1 \urcorner) \vee \neg True(\ulcorner L_1 \urcorner)$ is not classified as true in the Kripkean theories. Therefore, the problem is just inherent to Kripke's construction method basing on the Kleene valuation schemes, instead of on which language we are working with. The 3-vlaued logics are just so coarse that they cannot see the relation and differences among sentences having truth value $\frac{1}{2}$. In this case, when we assign $\frac{1}{2}$ to all of L_1, $\neg L_1$, and the Liar L at the initial stage, the logic will become blind to their differences — it cannot see that some sentences among them are the negation of another, while some other sentences are completely unrelated, so it will evaluate $L_1 = True(\ulcorner L_1 \urcorner) \vee \neg True(\ulcorner L_1 \urcorner)$ just the same as the sentence $True(\ulcorner L \urcorner) \vee True(\ulcorner L \urcorner)$. To solve this problem, one has to use a finer logic.

One way to have Kripke's theory recognise these classical tautologies is to use supervaluation instead of the Kleene logics. What remains the same is still to assign to

True both an extension X and an anti-extension Y with $X \cap Y = \emptyset$, and there could be sentences which are in neither of them. To evaluate the truth of a sentence ϕ, however, we do not simply use the extension and anti-extension based on some many-valued valuation scheme. Instead, we look at all candidate extensions Z of truth which are consistent with our evidence so far — $X \subseteq Z$ and $Z \cap Y = \emptyset$. Thus, Z contains every sentence which is already classified as true, and does not contain any sentence which is already classified as false. For each such Z, we have a classical model where the extension of the truth predicate is interpreted as Z. We can then evaluate ϕ in all these models. If ϕ is true in all of them, then we say ϕ has truth value 1, if ϕ is false in all of them, then we say ϕ has truth value 0, while if ϕ is true in some of them and false in others, then we say ϕ has truth value $\frac{1}{2}$:

Definition 3.11 (Supervaluation)**.** *The valuation scheme $V_{Sv} : (\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})) \to (Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\})$ is defined as follows: let $(X, Y) \in \mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$ be a pair of sets of sentences in $\mathcal{L}^+$, where $X \cap Y = \emptyset$. Then we define $v = V_{Sv}(X, Y) : Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\}$ as follows: let $\phi \in Sent(\mathcal{L}_S)$,*

1. *If for all Z such that $X \subseteq Z$ and $Z \cap Y = \emptyset$, $(\mathbb{N}, Z) \models \phi$, then $v(\phi) = 1$;*
2. *If for all Z such that $X \subseteq Z$ and $Z \cap Y = \emptyset$, $(\mathbb{N}, Z) \not\models \phi$, then $v(\phi) = 0$;*
3. *Otherwise, $v(\phi) = \frac{1}{2}$.*

The exact same definition for Kripke-jump (Definition 3.3) and the Kripkean extension of truth (Theorem 3.4) works for the supervaluation scheme.

This solves the problem for the Kripkean theories based on the Kleene logics:

Example 3.12. *Let $L_1 = True(\ulcorner L_1 \urcorner) \vee \neg True(\ulcorner L_1 \urcorner) \in Sent(\mathcal{L}_S)$. Then for any extension Z of True, we have $(\mathbb{N}, Z) \models True(\ulcorner L_1 \urcorner) \vee \neg True(\ulcorner L_1 \urcorner)$ because either $\ulcorner L_1 \urcorner \in Z$ or $\ulcorner L_1 \urcorner \notin Z$. Hence, L_1 will be added to the extension of truth at the first stage and will remain there throughout the whole process. Therefore, L_1 is classified as true in the supervaluation scheme.*

The supervaluation scheme can be developed in the infinitary propositional language as follows:

Definition 3.13 (Supervaluation)**.** *The valuation scheme $V_{Sv} : S \times S \to (Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\})$ is defined as follows: let $(S_1, S_2) \in S \times S$ be a pair of sets of sentences in S. Then we define $v = V_{Sv}(S_1, S_2) : Sent(\mathcal{L}_S) \to \{0, \frac{1}{2}, 1\}$ as follows: let $\phi \in Sent(\mathcal{L}_S)$,*

1. *If for all classical valuation $w : S \to \{0, 1\}$ such that $S_1 \subseteq \{t \in S \mid w(t) = 1\}$ and $\{t \in S \mid w(t) = 1\} \cap S_2 = \emptyset$, $[\![\phi]\!]_w = 1$, then $v(\phi) = 1$;*
2. *If for all classical valuation $w : S \to \{0, 1\}$ such that $S_1 \subseteq \{t \in S \mid w(t) = 1\}$ and $\{t \in S \mid w(t) = 1\} \cap S_2 = \emptyset$, $[\![\phi]\!]_w = 0$, then $v(\phi) = 0$;*

3. Otherwise, $v(\phi) = \frac{1}{2}$.

That is, if ϕ comes out true under every classical valuation consistent with our evidence so far, we add ϕ to our extension of truth; if ϕ comes out false under every such valuation, we add ϕ to our anti-extension; otherwise we leave ϕ undecided.

The Kripke-jump and the Kripkean extension of truth based on supervaluation are defined in the same way as the ones above (Definition 3.9 and Theorem 3.10). Let us denote the Kripkean extension of truth based on supervaluation as T_K^{Sv}.

We have the counterpart of Example 3.12 in the infinitary propositional language:

Example 3.14. *Let* $s_1 \in S$ *be such that* $d(s_1) = s_1 \vee \neg s_1$. *Then for classical valuation* $w : S \to \{0, 1\}$, *we have* $[\![s_1 \vee \neg s_1]\!]_w = 1$. *Hence,* s_1 *will be added to the extension of truth at the first stage and will remain there throughout the whole process. Therefore,* s_1 *is classified as true in the supervaluation scheme.*

3.4 Comparison with Kripke's Theory of Truth

We now discuss the relation between the Kripkean theories and the theory proposed in this thesis[3]. We will first formally compare the extensions of truth in the various theories, and then discuss the philosophical significance of the differences. In particular, I will argue that the proposed theory handles what I will call the "ungrounded tautologies" better than the Kripkean theories.

Recall that the extension of truth proposed in this essay is denoted by T, and for convenience we denote the extension of false sentences as F.

Lemma 3.15. *1.* $T_K^W \subsetneq T$.

2. T_K^S *and* T *are incomparable.*

Proof. 1. We prove the first part by proving $(X_\alpha, Y_\alpha) \subseteq (T, F)$ for all α. The base case is trivial. Now suppose $(X_\alpha, Y_\alpha) \subseteq (T, F)$ for some α. We need to show $(X_{\alpha+1}, Y_{\alpha+1}) \subseteq (T, F)$. By the definition of the Kripke-jump, we have $X_{\alpha+1} = \{s \in S \mid V(X_\alpha, Y_\alpha)(d(s)) = 1\}$. Let $s \in X_{\alpha+1}$ be arbitrary. We must have $D(s) \subseteq X_\alpha \cup Y_\alpha$ — otherwise, there is $t \in D(s)$ where $V(X_\alpha, Y_\alpha) = \frac{1}{2}$. Then by the weak Kleene logic, we must have $V(X_\alpha, Y_\alpha)(d(s)) = \frac{1}{2}$, contradiction. Then since $(X_\alpha, Y_\alpha) \subseteq (T, F)$, we have $D(s) \subseteq T \cup F$. As s only depends on true or false sentences, by Lemma 2.31 we have s itself is also either true or false. Since $s \in X_{\alpha+1}$, we have $V(X)(X_\alpha, Y_\alpha)(d(s)) = 1$. By inductive hypothesis, $X_\alpha \subseteq T$ and $Y_\alpha \subseteq F$. Therefore, since the valuation in this language is identical to weak Kleene logic when all propositional letter receive a classical value, by Lemma 3.8 we must have

[3]Recall Definition 2.26, the classification criteria of the proposed theory.

$[\![d(s)]\!]_v = 1$ for a valuation v where $v(t) = 1$ for all $t \in X_\alpha$ and $v(t) = 0$ for all $t \in Y_\alpha$. Therefore, $s \in T$. The proof for $Y_{\alpha+1}$ is similar. The limit case is straightforward.

To see that this is a strict inclusion, we consider two important examples.

(a) Let $s \in S$ be such that $d(s) = s \vee \neg s$. Then, as we discussed in the previous section, s is in neither T_K^W nor T_K^S. We show that it is in T. Clearly, $V_s = \{s\}$ and the ascription function $f_s : (0,1) \to (0,1)$ is such that $f_s(x) = 1$. Therefore, s has a unique fixed point $(1,1)$, so $s \in T$.

(b) Let $s_1, s_2 \in S$ be such that $d(s_1) = s_2$ and $d(s_2) = s_1 \vee \neg s_2$. Then according to the weak Kleene version of Kripke's construction s_1 and s_2 receive value $\frac{1}{2}$ at the initial stage, and then since they only depend on each other, they will be assigned the same value throughout the whole process and will never enter the extension of truth. Therefore, $s_1, s_2 \notin T_K^W$. However, one can easily check that $f_{s_2}(x,y) = (max(1-x,y), y)$ has a unique fixed point $(1,1)$, and $f_{s_1}(x,y) = (x, max(x, 1-y))$ also has a unique fixed point $(1,1)$. Therefore, $s_1, s_2 \in T$.

2. We now show that T_K^S and T are incomparable. The two examples above also show that $T \not\subseteq T_K^S$. Now let s_3, s_4 be such that $d(s_3) = \neg s_3$ and $d(s_4) = s_3 \vee \top$. s_3 is the Liar, which is not true in either theory. However, s_4 becomes true at the second stage of the Kripke-jump based on the strong Kleene logic, and hence $s_4 \in T_K^S$. Nonetheless, s_4 is not true in the theory proposed in this thesis because it depends on a paradoxical sentence. Therefore, $T_K^S \not\subseteq T$.

□

Moreover, (T, F) is also a fixed point of the Kripke-jump based on the weak Kleene logic. This is an interesting result since it gives an extension in the Kripkean hierarchy of fixed points other than the minimal one.

Lemma 3.16. *(T, F) is a fixed point of the Kripke-jump based on the weak Kleene logic.*

Proof. We need to show that $(T, F) = \mathcal{J}_{WK}(T, F)$. Let $X = \{s \in S \mid V_{WK}(T, F)(d(s)) = 1\}$.

1. $X \subseteq T$: Let $s \in X$ be arbitrary. Since we are using weak Kleene logic, as in the proof above, We must have $D(s) \subseteq T \cup F$. As s only depends on true or false sentences, by Lemma 2.31 we have s itself is also either true or false and Since $s \in X$, we have $V_{WK}(T)(d(s)) = 1$. Therefore, $s \in T$. The proof for $\{s \in S \mid V_{WK}(T, F)(d(s)) = 1\}$ is similar.

2. $T \subseteq X$: Let $s \in T$ be arbitrary. We have $V_{WK}(T, F)(d(s)) = 1$. By the definition of the Kripke-jump, we have $X = \{s \in S \mid V_{WK}(T, F)(d(s)) = 1\}$. Therefore, $s \in X$.

□

Lemma 3.17. *T and T_K^{Sv} are incomparable.*

Proof. Consider again the pair of sentences $s_1, s_2 \in S$ with $d(s_1) = s_2$ and $d(s_2) = s_1 \vee \neg s_2$. We have seen that $s_1, s_2 \in T$. However, we show that they are not in the Kripkean extension of truth based on supervaluation. We first show that for all α, we have $s_1, s_2 \notin X_\alpha \cup Y_\alpha$. The base step is trivial. Now suppose $s_1, s_2 \notin X_\alpha \cup Y_\alpha$ for some α. Then let $w_1 : S \to \{0, 1\}$ be a classical valuation such that $w_1(s_1) = 0$, $w_1(s_2) = 1$, $X_\alpha \subseteq \{t \in S \mid w_1(t) = 1\}$, and $\{t \in S \mid w(t) = 1\} \cap Y_\alpha = \emptyset$. Let $w_2 : S \to \{0, 1\}$ be a classical valuation such that $w_2(s_1) = 1$, $w_2(s_2) = 0$, $X_\alpha \subseteq \{t \in S \mid w_2(t) = 1\}$, and $\{t \in S \mid w(t) = 1\} \cap Y_\alpha = \emptyset$. There exists such w_1, w_2 because $s_1, s_2 \notin X_\alpha \cup Y_\alpha$. Then we have $[\![d(s_1)]\!]_{w_1} = 1$, $[\![d(s_1)]\!]_{w_2} = 0$, $[\![d(s_2)]\!]_{w_1} = 0$, and $[\![d(s_2)]\!]_{w_2} = 1$. Therefore, $s_1, s_2 \notin X_{\alpha+1} \cup Y_{\alpha+1}$. The limit case is standard. By induction, we have $s_1, s_2 \notin X_\alpha \cup Y_\alpha$ for all α. Then, in particular, $s_1, s_2 \notin T_K^{Sv}$.

On the other hand, let $d(s_1) = \neg s_1$, $d(s_2) = s_1 \vee \neg s_1$, then $s_2 \in T_K^{Sv}$ but $s_2 \notin T$.

□

We now summarise the differences between these theories and discuss their philosophical significance.

As we saw in the above proofs, there are examples where the proposed theory classifies as true, but the Kripkean theories based on Kleene logics do not: the s such that $d(s) = s \vee \neg s$ and s_1, s_2 with $d(s_1) = s_2$ and $d(s_2) = s_1 \vee \neg s_2$. The supervaluation version of Kripke's theory classifies s as true, but s_1 and s_2 are not true even in this version.

Thus, there are two important questions to ask about the differences between the theories:

1. Are there any reasons we should think s_1, s_2 are true?

2. Are there any reasons we should distinguish between s and the pair s_1, s_2, with regard to whether they are true?

In the other direction, there are also sentences classified as true by the Kripkean theory based on the strong Kleene logic and the supervaluation version, but not by the theory proposed in this paper: the s_3, s_4, s_5 such that $d(s_3) = \neg s_3$, $d(s_4) = s_3 \vee \top$, and $d(s_5) = s_3 \vee \neg s_3$. Then s_4 is classified as true in both Kripkean theories, s_5 is classified as true in the supervaluation version only, while neither of them is classified as true in the proposed theory. Therefore, although s and s_5 both have the form of a classical tautology ($p \vee \neg p$), the proposed theory classifies one of them as true and the other as not true, while all the Kripkean theories do not distinguish between them with regard to truth — those based on the Kleene logics classify both of them as not true, while the supervaluation version classifies both of them as true. Thus, we would also want to ask:

3. Are there any reasons we should distinguish between s and s_5 with regard to whether they are true?

Before answering these questions, let us first see that these are also problems in the original Kripkean theories in the first order language. The sentence $L = True(\ulcorner L \urcorner) \vee \neg True(\ulcorner L \urcorner)$ is not true in the Kripkean theories based on the Kleene logics, while it is true in the supervaluation version; while $L_1 = True(\ulcorner L_2 \urcorner)$ and $L_2 = True(\ulcorner L_1 \urcorner) \vee \neg True(\ulcorner L_2 \urcorner)$ are not classified as true in all the Kripkean theories. Therefore, these phenomena are inherent to the Kripkean theories instead of my formulation of them in the infinitary propositional language.

I will now argue that the answer to all three questions above is positive, and hence the proposed theory gives a finer classification of truth in these aspects.

For the first two questions, let us look at the corresponding sentences of s_1 and s_2 in the natural language. s_1 corresponds to the sentence L_1 = "L_2 is true" and s_2 to the sentence L_2 = "L_1 is true or L_2 is not true". As the theory I proposed shows, there is no contradiction with T-schema if we assume that both sentences are true. Moreover, this is the only configuration of truth assignment that does not yield a contradiction.

Furthermore, there is a sense in which this pair of sentences are just the same as the sentence $L = True(\ulcorner L \urcorner) \vee \neg True(\ulcorner L \urcorner)$, which is the corresponding sentence of s — if we substitute what L_1 says into L_2, then L_2 really just expresses the tautology "L_2 is true or L_2 is not true". The only difference is that not only are the reference patterns of L_1 and L_2 circular, but the tautological structure is also embedded in the circularity. Let us call these sentences "ungrounded tautologies". The Kripkean theories based on the Kleene logics recognise none of the ungrounded tautologies as true, because in these theories, a true sentence has to be grounded, even if they are just harmless tautologies. The supervaluation version of Kripke's theory seems to be an improvement, as it recognises at least the self-referential tautology s as true. Nevertheless, it has a more serious problem of making an arbitrary distinction between s and the pair s_1, s_2. Note that this is a very general problem — given any tautology ϕ, we can use this trick to replace a subformula ψ of the tautologies by some θ saying "ψ is true". Then the new sentence would not be recognised as true by the Kripkean theories.

The reason for this phenomenon is that the minimal fixed points of Kripke's theories — regardless of the versions — are essentially built upon the idea of groundedness. One starts with the empty set, and then applying the Kripke-jump gives us sentences that can be seen as true grounded on the truth of nothing (so they are simply grounded), and then we keep applying the Kripke-jump to get more and more sentences that can be seen as true grounded on the truth of sentences that we have already seen as grounded. For a sentence to be considered innocuous, it has to find some sentences that are more innocuous than itself to be grounded on. In the Kripkean theory based on the Kleene logics, the s with $d(s) = s \vee \neg s$ is not in the minimal fixed point because the only sentence

that it can seek help from is itself, so it can never be proved innocent. The supervaluation version resolves this problem, basically by taking additional care of sentences that have an explicit form of a classical tautology. However, as we have seen, it only needs a level of disguise to make a tautology look suspicious again: the supervaluation version only saves circular sentences that have explicit forms of classical tautologies, but when the tautological structures are also embedded in the circularity, it will get lost again.

Lastly, let us resolve the third question — does the theory I propose make an arbitrary distinction between s and s_5? Recall that s corresponds to the sentence $L = True(\ulcorner L \urcorner) \vee \neg True(\ulcorner L \urcorner)$, while s_5 corresponds to the sentence $L_5 = True(\ulcorner L_3 \urcorner) \vee \neg True(\ulcorner L_3 \urcorner)$, where L_3 is the Liar. It seems that both are just tautologies of the form $p \vee \neg p$, and it is only that the p in L_5 is replaced by the Liar.

Firstly, there is a trivial response — since the Liar is well recognised as a paradox, one has enough reason to be cautious and make a distinction between any sentences involving the Liar and those that do not.

However, there is an even more substantial reason to make a distinction between L and L_5 with respect to whether they are true. Under some very weak assumptions about the truth predicate, claiming L_5 to be true is equivalent to claiming $True(\ulcorner L_3 \urcorner) \vee True(\ulcorner L_3 \urcorner)$ to be true — which is, I believe, not what one would want to do.

Lemma 3.18. *Under some very weak assumptions about the truth predicate, we have* $True(\ulcorner L_5 \urcorner) = True(\ulcorner True(\ulcorner L_3 \urcorner) \vee \neg True(\ulcorner L_3 \urcorner) \urcorner)$ *is equivalent to* $True(\ulcorner True(\ulcorner L_3 \urcorner) \vee True(\ulcorner L_3 \urcorner) \urcorner)$.

Proof.
1. $True(\ulcorner True(\ulcorner L_3 \urcorner) \vee \neg True(\ulcorner L_3 \urcorner) \urcorner)$, assumption;
2. $True(\ulcorner True(\ulcorner L_3 \urcorner) \vee L_3 \urcorner)$, by definition we have $L_3 = \neg True(\ulcorner L_3 \urcorner)$;
3. $True(\ulcorner True(\ulcorner True(\ulcorner L_3 \urcorner) \urcorner \vee L_3 \urcorner))$, assume $True$ is idempotent;
4. $True(\ulcorner True(\ulcorner True(\ulcorner L_3 \urcorner) \urcorner) \vee True(\ulcorner L_3 \urcorner) \urcorner)$, assume $True$ is distributive over $\vee$;
5. $True(\ulcorner True(\ulcorner L_3 \urcorner) \urcorner) \vee True(\ulcorner True(\ulcorner L_3 \urcorner) \urcorner)$, assume $True$ is idempotent.

Moreover, with these assumptions, all steps are reversible.

□

Therefore, when L_3 is the Liar sentence, if one wants to classify $True(\ulcorner L_3 \urcorner) \vee \neg True(\ulcorner L_3 \urcorner)$ as true, one must either classify $True(\ulcorner L_3 \urcorner) \vee True(\ulcorner L_3 \urcorner)$ as well, or one must give up some assumptions above. It is very unclear which choice one has to make, and this fact suggests that there is a fundamental tension between any sentence involving the Liar with our classical intuition about truth — not just the tension between the Liar itself and Tarski's T-schema. Therefore, there is sufficient reason to treat any sentence involving the Liar with caution, even if it has the form of a classical tautology. Thus, it is not arbitrary — or even advisable — to distinguish between s and s_5.

There are further interesting questions about the differences between the Kripkean theories and the theory proposed in this paper.

Firstly, one might wonder are the ungrounded tautologies[4] the only kind of sentences that the proposed theory classifies as true, but the Kripkean theories do not? If not, what are some other examples?

Moreover, besides the minimal fixed points, Kripke also proposed other fixed points of the Kripke-jump which could have meaningful interpretations. One fixed point that seems to be closely related to the theory developed here is the "largest intrinsic fixed point", which he characterises as "the unique 'largest' interpretation of $T(x)$ which is consistent with our intuitive idea of truth and makes no arbitrary choices in truth assignments." (Kripke, 1975: p. 709 - 710). It would be interesting to compare the theory proposed in this paper with this fixed point.

Another interesting question arises from the fact that while $T_K^W \subseteq T$, we find that T_K^S and T are incomparable. In some sense, this is to be expected — after all, the definition of hereditarily successful really reflects the weak Kleene intuition that if a part of a sentence is neither true nor false, then the whole sentence is neither true nor false. It would be interesting to build up a definition of truth from the successful sentences in a way closer to the strong Kleene intuition that, for example, for a disjunct to be true we only need one of its disjuncts to be true. Then we might be able to get a theory of truth that contains the Kripkean theory based on the strong Kleene logic.

To conclude this section, I want to emphasise an essential difference between Kripke's theory and the proposed one. In Kripke's theories, one always starts from the bottom, and all different candidates for an extension of truth (whether it is the minimal, intrinsic, or the largest intrinsic fixed point, or even a fixed point based on some none-empty sets) are eligible because they are all fixed points of the Kripke-jump — which means that they are grounded on something. However, in the proposed theory, no importance is given to groundedness. The only thing that matters is whether there is a unique way to assign a truth value to a sentence. I believe this is a more natural way to think about truth, and it is how people actually approach potentially problematic sentences in natural language.

[4]To answer these questions, one of course needs to define precisely the notion of "ungrounded tautologies".

Chapter 4

Comparison with the Revision Theory of Truth

Besides Kripke's theory, there is another important theory of truth that keeps the truth predicate non-hierarchical — the Revision Theory of truth. In fact, my own proposal in this thesis is a modification of this theory. In this section, I will first introduce the Revision Theory, and present its explanation of the Liar paradox. I will then develop the theory in the infinite propositional language and discuss my objections. Lastly, I will elucidate the similarities and differences between the Revision Theory and the theory proposed in this thesis.

4.1 Introduction to the Revision Theory of Truth

The Revision Theory of truth builds on a theory of definition that allows circularity. The essential feature of this theory is that it does not give a fixed extension to the predicate $True$, so one can only talk about whether a sentence is "true" relative to a "stage" of evaluation.

The idea is that in a non-circular definition of a predicate (e.g. "is an even number"), one can determine its extension from its defining rule (e.g. "is divisible by two"). However, this is impossible when the defining rule also mentions the predicate itself, because then to apply the rule one has to know its extension in the first place. For a trivial example, if I define a number to be "good" if the number is good, then I cannot determine the extension of "good" because I need to know it before applying this definition. In this case, revision theorists argue that the definition provides "a rule that can be used to calculate what the extension should be once we make a hypothesis concerning the extension" (Gupta and Belnap, 1993: p119). Given the hypothesis for the extension, together with all the

This Chapter is based on an essay I wrote for an individual project done with Dr. Thomas Schindler.

relevant facts in our world, one can calculate a revised extension, which is then used as another hypothesis, and the process goes on. In particular, they suggest that Tarski's T-schema — "ϕ is true if and only if ϕ" — can be used as a definition of the predicate $True$.

For example, suppose that there are two sentences in our language γ = "γ is not true" and δ = "grass is green" and we apply Revision Theory to calculate the extension of $True$ at each stage. The relevant fact in our world here is that grass is indeed green. Then the T-schema yields two partial definitions for $True$: (i) "γ is true $=_{def}$ γ is not true" (ii) "δ is true $=_{def}$ grass is green" (note that δ, or "grass is green", on the left-hand side is just a sentence — a syntactical object — in the language, while the right-hand side of this partial definition is a proposition regarding a fact in our world). These are partial definitions for updating the hypothetical extension of $True$.

Let us first make the hypothesis that the extension of $True$ is the empty set. At stage 0, according to the hypothesis and the relevant fact (i.e. grass is indeed green), the criterion "γ is not true" and "grass is green" are both satisfied, and hence one revises the hypothesis to include both γ and δ in the extension of $True$ at stage 1. However, then "γ is not true" is not satisfied although nothing changes the fact that grass is green, so one revises the hypothesis again to conclude that the extension of $True$ only contains δ at stage 2. This process goes on indefinitely, and one can see that it is impossible to give a stabilised extension for the predicate $True$ regardless of what the hypothesis one makes at stage 0 is — the extension of $True$ always includes δ after stage 1 but oscillates between including and not including γ at each stage. Thus, we can only talk about whether a sentence is "true" relative to a stage of evaluation.

Then the reaction we have towards the Liar is reflected by the sequence of revision of the extension of $True$: assuming it to be true at first would lead one to conclude later that it is false, and vice versa. The revision theorists claim that the non-paradoxical sentences (e.g. "grass is green") are those that are always included in the extension of $True$ after some finite stage regardless of the hypothesis at stage 0, or always excluded in the extension of $True$ regardless of the initial hypothesis (Gupta and Belnap, 1993: p137).[1]

4.2 Revision Theory in the Infinitary Propositional Language

I now develop the theory in the infinitary propositional language, based on Field (2008: pp. 186 - 187). Recall that in this language, all relevant facts about the world are decided

[1]The Truth Teller sentence "This sentence is true" is thus classified as paradoxical because although it stabilises to either true or not true after some finite stage, exactly which one depends on the hypothesis at stage 0.

in the metatheory and are represented as $\top$ or $\bot$ in the language. Therefore, in order to define the revision rule, we only need to consider the hypothetical extension of truth. Moreover, in this language, there is no predicate. However, we can just use the valuations: extensions of truth can be seen as the set of sentences that are assigned 1 by a valuation.

Definition 4.1 (Revision Rule)**.** *The* revision rule $\tau : \mathcal{P}(S) \to \mathcal{P}(S)$ *is defined as follows. Let* $S_1 \subseteq S$ *and we define* $\tau(S_1)$*. Let* $v : S \to \{0, 1\}$ *be the valuation such that* $v(s) = 1$ *if* $s \in S_1$ *and* $v(s) = 0$ *if* $s \notin S_1$*. Then we define* $\tau(S_1) = \{s \in S \mid v(d(s)) = 1\}$*.*

That is, given a hypothetical extension of truth S_1, the revision rule gives an updated hypothesis $\tau(S_1)$, containing all those s such that $d(s)$ is true under the valuation which assigns 1 to every propositional letter in S_1 and 0 to the other letters. Keep applying the revision rule gives us the revision sequence:

Definition 4.2 (Revision Sequence)**.** *A* revision sequence *is a transfinite sequence of length ordinal* $S_\alpha \subseteq S$ *such that* $S_{\alpha+1} = \tau(S_\alpha)$ *for all* α *and for a limit ordinal* λ*,* S_λ *satisfies* $\{s \mid (\exists\beta < \lambda)\forall\delta(\beta \leq \delta < \lambda \to s \in S_\delta)\} \subseteq S_\lambda \subseteq \{s \mid (\forall\beta < \lambda)\exists\delta(\beta \leq \delta < \lambda \wedge s \in S_\delta)\}$*.*

Remark 4.3. *At the limit stage* λ*, a sentence is included if there is some stage* $\beta < \lambda$ *such that for all later stages* $\delta \geq \beta$*, the sentence is included in the extension of truth; and it is excluded if for all stages* $\beta < \lambda$ *there is some later stage* $\delta \geq \beta$ *such that the sentence is not included in the extension of truth.*

In other words, a sentence is included at the limit stage if it has been stably included in the extension of truth after some sufficiently large stage, and it is excluded otherwise.

Then a theory of truth is given by including sentences that are stably included in the extension of any revision sequence after a sufficiently large stage:

Definition 4.4 (Revision Truth)**.** *A sentence* $s \in S$ *is* true *if for all revision sequence* S_α *there is* γ *such that for all* $\beta \geq \gamma$*,* $s \in S_\beta$*.*

Example 4.5. *1. Liar: let* $s \in S$ *be such that* $d(s) = \neg s$*. Then* s *always oscillates between being true and false in any revision sequence.*

2. Truth Teller: let $s \in S$ *be such that* $d(s) = s$*. Then* s *is true in a revision sequence* S_α *if and only if* $s \in S_0$*.*

Therefore, none of them is true in the Revision Theory.

4.3 Issues with the Revision Theory

I will now argue that this theory is not satisfactory in two ways:

1. The explanation it provides for the truth does not work well for non-paradoxical sentences.

2. It overemphasises the significance of the revision sequences.

Moreover, in the discussion of the second point, I will illustrate how the theory proposed in this thesis resolves the problems.

Firstly, by using the Revision Theory for both paradoxical and non-paradoxical sentences, it provides explanations for the former at the expense of the latter. Since there is no fixed extension for $True$, revision theorists claim that the semantic status for a sentence is the pattern of its revision sequence (Herzberger, 1982a: p492) or the "signification of truth is a rule of revision"(Gupta and Belnap, 1993: p139). However, this would yield a very unnatural explanation for non-paradoxical sentences. For example, the semantic status for "grass is green" is the pattern of revision sequence where the sentence is always included in the extension of $True$, thus, according to the explanation of the Liar, our reaction towards it should be as follows: at some initial stage we make a hypothesis that it is "true" or "false", then we revise our hypothesis according to the relevant fact in our world, and we continue doing it indefinitely (or maybe by some meta-theorem we can ensure ourselves that it will always stay in the extension). However, actually we just say it is true and that's it — we do not keep revising our conclusion.

The second objection to Revision Theory is that the theory overstates the importance of the revision sequence. This is because there are sentences that keep oscillating for all initial hypotheses except for one where it stabilises.

Consider the two sentences below:

1. $\phi = (\neg True(\ulcorner\phi\urcorner) \wedge \neg True(\ulcorner\psi\urcorner)) \vee (True(\ulcorner\phi\urcorner) \wedge True(\ulcorner\psi\urcorner))$

2. $\psi = (True(\ulcorner\phi\urcorner) \wedge \neg True(\ulcorner\psi\urcorner)) \vee (True(\ulcorner\phi\urcorner) \wedge True(\ulcorner\psi\urcorner))$

Are they paradoxical? Intuitively, we would argue — in the same way when we face the Liar — as follows: assume ϕ is true and ψ is false, then we can calculate from the T-schema that ϕ is false and ψ is true, a contradiction. Then we check all the other possibilities, and find that all of them lead to a contradiction, except for the case when we assume both of them to be true. Therefore, I believe we should conclude that they are not paradoxical — this is like a logic puzzle and both sentences are true. However, the Revision Theory[2] classifies them as paradoxical, because they do not stabilise on all initial hypotheses. In a language that contains these two sentences, one can calculate that the revision sequence for ϕ and ψ is as follows:

ϕ	1	0	0	1	...
ψ	0	1	0	0	...

[2]Or, the most natural extension of this theory for sentences that refer to each other. The sketch for this theory in Chapter 4 of Gupta and Belnap 1993 is discussed only in the presence of a directly self-referential sentence, without considering sentences that refer to each other.

Starting from any initial hypothesis for the extension of $True$ that does not include both of them, one would end up in the above circle and conclude that the extension of $True$ keeps oscillating between including and not including ϕ and ψ. However, if one includes both in the initial hypothesis, then the extension of $True$ will always include both of them after the first stage.

The point of this example is that even if there is a hypothesis that stabilises, making a random initial guess may not lead to it no matter how many times one revises her hypothesis. I believe what is really important is whether there is a hypothesis that stabilises. This is exactly how the proposed theory works — checking whether there is a unique fixed point for the ascription function, instead of whether all initial hypothesis stabilises to the same point after applying the function for sufficiently many times.

The revision sequence is like Newton's algorithm for finding the root of a function: it makes an initial guess and revises it according to a certain update function, and the output gets closer and closer to the root. However, for the algorithm to work, one needs the function to be "nice" and the initial guess should be "close enough" to the root. In the above example, it turns out that one is "close enough" to the root only when one starts with the root.

We can formally see this in the propositional language:

Example 4.6. *Let* $d(s_0) = (\neg s_0 \wedge \neg s_1) \vee (s_0 \wedge s_1)$ *and* $d(s_1) = (s_0 \wedge \neg s_1) \vee (s_0 \wedge s_1)$*. Then* s_0 *and* s_1 *represent the two sentences in the above example. They are not true according to the Revision Theory (Definition 4.4). However, the ascription function* f_{s_0} *and* f_{s_1} *are both successful, both having the unique fixed point* $(1, 1)$*. Therefore, both sentences are true according to the theory proposed in this thesis.*

This shows the first difference between the Revision Theory and the theory proposed in this thesis: the former emphasises the revision sequence resulting from applying an update rule repeatedly, while the latter focuses on whether there is a unique fixed point for the rule. We will now see a second essential difference between the two theories by considering an attempt to resolve the issue mentioned above.

The most straightforward way to incorporate cases like Example 4.6 is to say that if there is a unique initial hypothesis for $True$ that stabilises, then the sentences are non-paradoxical. Let us call sentences satisfying this modification *unproblematic*:

Definition 4.7. *A sentence* $s \in S$ *is* unproblematic *if for all revision sequence* S_α *there is* γ *such that for all* $\beta \geq \gamma$*,* $s \in S_\beta$*, or if there is a unique initial hypothesis* S_0 *such that for all revision sequence starting with* S_0*, there is* γ *such that for all* $\beta \geq \gamma$*,* $s \in S_\beta$*.*

However, this definition fails to resolve the above issue. In fact, it creates an even worse problem — the semantic status of a sentence will depend on which sentence system one works with.

If we work in a sentence system (S, d) where the s_0 and s_1 in Example 4.6 are the only sentences in S. Then both of them are unproblematic, because $\{s_0, s_1\}$ is the only initial hypothesis that gives rise to a revision sequence where the two sentences eventually stablise. However, let (S', d') be an extension of (S, d) where we add the Truth Teller s_2 with $d'(s_2) = s_2$. Then both $\{s_0, s_1\}$ and $\{s_0, s_1, s_2\}$ are initial hypotheses that stabilises, so s_0 and s_1 are not unproblematic in (S', d').

Formally, we have just seen:

Lemma 4.8. *There are sentence systems (S_1, d_1) and (S_2, d_2), $s_1 \in S_1, s_2 \in S_2$ such that s_1 is isomorphic to s_2, s_1 is unproblematic in S_1 but s_2 is not so in S_2.*

This issue arises because the revision rule is defined globally for the predicate $True$ — to apply the rule, one has to give an initial hypothesis for every sentence in the language, and then the revision rule is applied to all sentences at once. Therefore, it gives a chance for a sentence to influence the semantic status of another completely irrelevant sentence — in the example above, s_0 and s_1 are sentences referring to each other, while having no relation to the Truth Teller s_2, but the addition of s_2 changes the semantic status of s_0 and s_1.

Admittedly, it is quite possible that further modifications would give a version of Revision Theory that solves these problems, but I think they would be a bit *ad hoc* in the sense that less and less importance would be given to the revision pattern, contrary to the elegance of this idea when only directly self-referential sentences are present. Analogies with Newton's algorithm continue here: I believe the necessity for these modifications results from putting too much weight on a technical tool for finding the root. It is whether the function has a root that is important, instead of the revision sequence that sometimes leads to the root.

Chapter 5

Other Aspects of the Theory

In this chapter, we discuss how the theory developed in this thesis can be used to shed light on other philosophical questions about truth and paradox. We will first see a way that the hypodoxical sentences — like the Truth Teller — can also be seen as paradoxical. Then, I will use the theory to formally capture the intuition that the Liar circle can be reduced to the Liar sentence itself. Lastly, I will discuss how one might modify the theory to respect certain intuitions about sentences involving the Liar.

5.1 Paradoxical Hypodox

In the literature, there are arguments implying that certain hypodoxical sentences — like the No-No paradox — are paradoxical (Sorensen, 2001: Ch11; Cook, 2011). However, these arguments either rely on contentious metatheoretical assumptions or on strong principles governing the truth predicate. In this section, I will first present two arguments supposedly demonstrating that the No-No paradox is a genuine paradox. I will then argue that each of them depends on unwarranted assumptions. Finally, using the technical tools developed in this thesis, I will offer a new argument that every hypodox can indeed be treated as a paradox, albeit in a weaker sense than the sense in which the Liar is paradoxical.

The No-No paradox is the pair of sentences L_1: "L_2 is not true." and L_2: "L_1 is not true." It is easily seen that assigning one of them as true and the other as false are the only two possibilities that do not lead to a contradiction. However, we can not decide which one is true and which one is false, so, like the Truth Teller, they are hypodoxical sentences.

Cook (2011) presents two arguments for the No-No paradox being a genuine paradox[1].

[1]I will outline both arguments below. However, note that Cook (2011) only endorses the second one. He presents the first argument only because he claims that Sorensen (2001) draws from it the conclusion that the symmetry principle should be abandoned. Cook himself uses the second argument to establish symmetry as a mathematical theorem rather than a philosophical principle.

The first one appeals to a metatheoretical principle involving symmetry. Noticing that the two sentences in the No-No paradox are completely symmetrical — each stating that the other is not true — the principle claims that "there seem to be no principled grounds for any semantic distinction between" (Cook, 2011: p. 468) the two sentences, and hence they must have the same truth value (Sorensen, 2001: p. 166). However, as we have already noted, they must have opposite truth values to avoid a contradiction. Therefore, the No-No paradox is really paradoxical.

The problem with this argument is to conclude that the two sentences must have the same truth value from their symmetrical structure. Of course, we should expect some symmetry of their semantic status, but having the same truth value is not the only way two sentences can have a symmetrical semantic status. In fact, the way they are hypodoxical is already symmetrical. Together with T-schema, there are two situations: if both are true or both are false, then they generate a contradiction; while if one of them is true and the other is false, then they are consistent with each other. This is already symmetrical — in any situation, we can swap the truth value of the two sentences while staying in the same situation. It is not that we have decided that L_1 must be true and L_2 must be false for them to be consistent, but rather any one of them can be true and the other false. They remain symmetrical in this respect.

This problem is most evident in a similar example in mathematics. Let x, y be two integers such that $x = -y$ and $y = -x$. Then they also have a symmetrical structure as do the two sentences in the No-No paradox. However, no one would conclude that x must be equal to y because of this symmetry, and hence the solution to this set of equations is $x = y = 0$. Rather, any pair $(n, -n)$ is a solution to these equations. Moreover, this solution set is symmetrical with respect to x and y — whenever $x = p$ and $y = q$ is a solution, $x = q$ and $y = p$ is also a solution. The same applies to the No-No paradox. Symmetrical structure only ever suggests that the two sentences play an interchangeable role, so that swapping their truth value should not change the situation we are in. It does not suggest that they must have the same truth value.

Therefore, the metatheoretical principle that symmetrical sentences must have the same truth value is unwarranted, and the first argument fails to show that the No-No paradox is a genuine paradox.

The second argument given by Cook is that under mild conditions on the Truth predicate, we can actually prove that the two sentences in the No-No paradox must receive the same truth value — "the symmetry principle is not a philosophically or intuitively motivated metatheoretic principle that can be abandoned in the face of recalcitrant data, but is instead an object language theorem" (Cook, 2011: p. 474). However, I will argue that the condition he gives is not mild at all — namely that the Truth predicate satisfies provability conditions and hence Löb's theorem: if $\vdash True(\ulcorner \phi \urcorner) \to \phi$, then $\vdash \phi$, i.e., if a sentence's truth implies this sentence, then this sentence holds.

There are two ways to see that this is a very implausible condition. First, let ϕ be "grass can fly". Then certainly one would agree that if "grass can fly" is true, then grass can fly, although no one would conclude thus that grass can fly. In fact, proving Löb's theorem from the provability conditions is exactly the same argument used in Curry's paradox to show that any sentence holds (Smith, 2013: p34), so one surely does not want to accept this. Second, as mentioned in (Smith, 2013) Löb's theorem implies Gödel's second incompleteness theorem (Kreisel, 1965), which in this context states that if Cook's theory of truth is consistent then "it is not true that $0 = 1$" does not hold. Such a truth theory is hardly interesting since it even advocates a radical judgement of truth on non-paradoxical sentences.

Hence, both arguments fail to show that the No-No paradox is a genuine paradox. Like the Truth Teller, they do not generate any contradiction on their own. Nevertheless, I will now argue — using the formal tools developed in this thesis — that there is a weaker sense in which they are indeed paradoxical. In fact, I will show that all hypodoxes are paradoxical in this sense.

Let us start by recalling the classification of the semantic status of sentences in Definition 2.26. We classified a sentence s according to the behaviour of its ascription function f_s — whether it has no fixed point (or, using our terminology in Definition 2.14, it is paradoxical), has multiple fixed points (it is hypodoxical), or it has a unique fixed point and f_t has a unique fixed point for all t that s depends on (it is hereditarily successful).

There is clearly another possibility of f_s that we did not cover — f_s has a unique fixed point, but f_t does not have a unique fixed point for some t that s depends on (in our terminology, f_s is successful but not hereditarily successful). What can we say about sentences having these kinds of ascription functions?

In fact, these sentences are exactly those depending on hypodoxical sentences:

Lemma 5.1. *Let $s \in S$ be a sentence in the sentence system (S, d). If f_s is successful, but not hereditarily successful, then there exists $t \in D_s$ such that f_t is hypodoxical.*

Proof. By Lemma 2.23, for all $t \in D_s$, f_t has a fixed point. Since f_s is not hereditarily successful, there exists $t \in D_s$ such that f_t is not successful, so it must have multiple fixed points. □

One might wonder whether this gives a good reason to believe that some hypodoxical sentences can be given a truth value after all: although one cannot decide its truth value by looking at itself, one might find another sentence which depends on it, but together they induce a function which has a unique fixed point. Then we can use this fixed point to determine the truth value of the hypodoxical sentence. However, this is not possible, since one can find sentences giving opposite truth values to the same hypodoxical sentence. Let us give an example.

Example 5.2. *1. Let $d(s_0) = s_0$ and $d(s_1) = (s_0 \wedge s_1) \vee (s_0 \wedge \neg s_1) \vee (\neg s_0 \wedge \neg s_1)$. s_0 is the Truth Teller, which is hypodoxical. However, f_{s_1} has the following table:*

(x, y)	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$f_{s_1}(x, y)$	$(1,1)$	$(0,1)$	$(1,1)$	$(0,1)$

We can see that it has a unique fixed point, which is $(1,1)$. Thus, according to s_1, both itself and the Truth Teller are true, since this is the only way to consistently assign truth values to them.

2. Let $d(s_0) = s_0$ again, while $d(s_2) = s_0 \wedge \neg s_2$. Then f_{s_2} has the following table:

(x, y)	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$f_{s_2}(x, y)$	$(0,1)$	$(0,0)$	$(1,1)$	$(0,0)$

We can see that it has a unique fixed point, which is $(0,0)$. Thus, according to s_2, both itself and the Truth Teller are false, since this is the only way to consistently assign truth values to them.

This example suggests a way to turn a hypodoxical sentence like the Truth Teller into a real paradox in natural language. Let L_0 be the Truth Teller, and let L_1 be the sentence "both L_0 and L_1 are true, or L_0 is true and L_1 is not true, or both L_0 and L_1 are not true". Let L_2 be the sentence "L_0 is true L_2 is not true". Note that L_i is represented by s_i in the above example. Now we have a paradox in the same way that the Liar is a paradox: assume L_0 is true. If L_2 is true, then according to what it says, L_2 is not true, which is a contradiction. If L_2 is not true, then according to what it says, L_0 is not true, which is also a contradiction. Therefore, L_0 cannot be true. Now, assume L_1 is true. Then one of its conjuncts is true. However, when L_0 and L_1 are both true, none of the conjuncts can be true. Therefore, L_1 cannot be true. However, then one of its conjuncts — "L_0 is true and L_1 is not true" — is, after all, true. This means that L_1 is true. This is a contradiction. Hence, we find a paradox.

One might wonder whether the above way of turning the Truth Teller into a paradox is trivial. It seems like when we are deriving a contradiction, we appeal to two sentences that the Truth Teller does not depend on, so the problem is not related to Truth Teller itself but with the added sentences. After all, given any sentence — paradoxical or not — one can add a paradox like the Liar and then derive a contradiction. Then, although it seems like these sentences create a paradox together, the problem is only with the added paradoxical sentence, not the original one. For example, consider this argument for turning "grass is green" into a paradox. Assume "grass is green" is true. If the Liar is true, then the Liar is not true, a contradiction. If the Liar is not true, then it is true, again a contradiction. Thus, "grass is green" cannot be true, but still, we can clearly

derive a contradiction by going through the Liar paradox again. Then, it seems like, by the same argument I gave above, I should say that this is a way of turning "grass is green" into a paradox.

This is not the case. There is an essential difference between this trivial reasoning and the one I gave above. The Liar sentence itself is paradoxical, so it is not surprising that it can be used to derive a contradiction. However, in the example I gave, L_1 and L_2 are not paradoxical on their own: as we saw, L_1 only depends on L_0 and itself, and there is a unique way of consistently assigning a truth value to L_1 and L_0; L_2 also only depends on L_0 and itself, and there is another unique way of consistently assigning a truth value to L_2 and L_0. The problem is only that when these three sentences are put together, they generate a paradox.

Admittedly, this still makes Truth Teller different from a paradox like the Liar, which is paradoxical on its own, but I believe this is an interesting phenomenon worth noting for the Truth Teller and provides a weak sense in which the sentence is paradoxical.

In fact, this observation generalises to all hypodoxical sentences. In the Truth Teller scenario, one finds two sentences that both depend on the Truth Teller (which is also the only sentence that the Truth Teller itself depends on) and yet force it to different truth values. In general, a hypodoxical sentence s can depend on sentences other than itself.[2] Its hypodoxicality arises because there is more than one way to assign truth values to the entire variable set[3] of s in a consistent manner. Therefore, identifying the whole variable set as the hypodox in the general case, one instead shows there is some sentence t in the hypodox for which we can find two sentences that depend on t and compel t to take conflicting truth assignments.

This phenomenon can be seen as a reason that the hypodoxes are really problematic — the truth value of some sentence in the hypodox can be fixed to any value, and together with other sentences in our language, they generate real paradoxes just like the Liar. We now formally prove this result.

We first need a useful lemma, which will let us find sentences that correspond to any ascription function.

Lemma 5.3. *Let $f : \{0,1\}^\alpha \to \{0,1\}^\alpha$ be an arbitrary function. There exists a sentence system (S,d) and $s \in S$ with $f_s = f$.*

Proof. In fact, it is the same as constructing a sentence in propositional language that has the truth table of an arbitrary logical connective. Let $V = \{s_0, s_1, \ldots, s_{\alpha'}\}$ be a set of propositional letters, where $s_i \neq s_j$ for all $i \neq j$ and $|\alpha' + 1| = |\alpha|$.

Let $P_0 : \{0,1\}^\alpha \to \{0,1\}$ be the projection map to the first coordinate. Let $X = \{\bar{x}_1, \ldots, \bar{x}\beta\} \subseteq \{T,F\}^\alpha$ be such that $P_0(f(\bar{x})) = T$ for all $\bar{x} \in X$. We write down the row

[2]In fact, up to isomorphism (as defined in Definition 2.39), the Truth Teller is the only hypodox that depends on itself.

[3]That is, s and the sentences that s depends on (recall Definition 2.10).

description σ_i for $\bar{x}_0$ using propositional letters in V with appropriate subscripts. For example, let $\alpha = 3$ and $\bar{x}_i = (T, F, T)$, then σ_i is the sentence $s_0 \wedge \neg s_1 \wedge s_2$. Then let $d(s) = \bigvee_{i=0}^{\beta} \sigma_i$. Note that other $d(s_i) \in V$ are determined in the same way by using the projection map to the $(i+1)$-*th* coordinate. One can check that $f_s = f$, where V is the variable set of s. □

Next, we show that for any hypodoxical sentence s, there are at least two ways of consistently assigning truth values to the variable set of s, so that some t in the variable set of s receives two different truth values. Note that this is not yet the result we are after — it only shows that there are two ways of assigning truth values to t which are contradictory to each other, not that there are two sentences that depend on t and force it to take these contradictory truth values. The proof is trivial, though it is important for the main result.

Lemma 5.4. *Let $s \in S$ be a hypodoxical sentence in the sentence system (S, d). Then there exists $t \in V_s$ such that it receives different truth values in two different fixed points: there exists $\bar{x}_1, \bar{x}_2$ of f_s such that $\bar{x}_1(\alpha(t)) \neq \bar{x}_2(\alpha(t))$. (Recall that $\alpha(t)$ is the index of t in the variable set V_s of s.)*

Proof. Since f_s has multiple fixed points, there exist $\bar{x}_1, \bar{x}_2$ such that $\bar{x}_1 \neq \bar{x}_2$. Then there must be some i such that $\bar{x}_1(i) \neq \bar{x}_2(i)$. Let t be the i-th sentence in the variable set of s. Then $\bar{x}_1(\alpha(t)) \neq \bar{x}_2(\alpha(t))$, as required. □

Finally, let us prove the main result of this section, which shows that for any hypodox, there are sentences that depend on some sentence t in the hypodox and force t to take contradictory truth values.

Theorem 5.5. *Let $s \in S$ be a hypodoxical sentence in the sentence system (S, d). Then there exist $s_1, s_2 \in S$ and $t \in V_s$ such that s_1 and s_2 depend on t, f_{s_1} and f_{s_2} are both successful, but the naive truth value [4] of t according to f_{s_1} is 0, while the naive truth value of t according to f_{s_2} is 1.*

Proof. Let $t \in V_s$ be such that there exist two fixed points $\bar{x}_1, \bar{x}_2$ of f_s such that $\bar{x}_1(\alpha(t)) \neq \bar{x}_2(\alpha(t))$, which exists by Lemma 5.4.

Let $\bar{x}_0, \ldots, \bar{x}_\alpha$ be the fixed points of $f_s : \{0,1\}^\alpha \to \{0,1\}^\alpha$. Define $\bar{y}_i$ such that $\bar{y}_i(0) = 0$, and $\bar{y}_i(j+1) = \bar{x}_i(j)$ for all $j < \alpha$; and $\bar{y}_i$ such that $\bar{z}_i(0) = 1$, and $\bar{y}_i(j+1) = \bar{x}_i(j)$ for all $j < \alpha$. That is, the 0-th coordinate of $\bar{y}_i$ is 0, while the rest is a copy of $\bar{x}_i$, and the 0-th coordinate of $\bar{z}_i$ is 1, while the rest is a copy of $\bar{x}_i$. Let $f_{s_1} : \{0,1\}^{\alpha+1} \to \{0,1\}^{\alpha+1}$ be a function such that $f_{s_1}(\bar{y}_0) = \bar{y}_0$, and $f_{s_1}(\bar{y}_i) = \bar{z}_i$ for all $i > 0$, and $f_{s_2} : \{0,1\}^{\alpha+1} \to \{0,1\}^{\alpha+1}$ be a function such that $f_{s_2}(\bar{z}_0) = \bar{z}_0$, and $f_{s_2}(\bar{z}_i) = \bar{y}_i$ for all $i > 0$. For any other $\bar{y} \in \{0,1\}^{\alpha+1}$, let $f_{s_1}(\bar{y}) = f_{s_2}(\bar{y}) = f_s(y_1, y_2, \ldots, y_{\alpha+1})$.

[4] Recall Definition 2.17.

By Lemma 5.3, we can find a sentence system (S, d) and $s_1, s_2 \in S$ such that $f_{s_1} = f_{s_1}$ and $f_{s_2} = f_{s_2}$. □

Thus, we have shown that every hypodox is paradoxical in a weak sense: we can always find sentences in the language which force some sentence in the hypodox to take contradictory truth values.

In fact, by the same proof method, one can show that for any given hypodox there exist sentences (depending on some member of that hypodox) whose ascription functions have arbitrarily chosen patterns of fixed points.

Lemma 5.6. *Let $s \in S$ be a hypodoxical sentence in the sentence system (S, d). Then there exist $s_1 \in S$ and $t \in V_s$ such that s_1 depends on t, and f_{s_1} is successful (paradoxical, hypothetical).*

There are several further questions worth exploring in this direction — namely, examining the detailed fixed point patterns of the ascription functions of hypodoxical sentences. For example, in Lemma 5.4, for any hypodoxical sentence s we only showed that there is a sentence $t \in V_s$ that can receive different truth values in different fixed points of f_s. We did not claim that s itself can also have different truth values. In fact, this is not the case:

Example 5.7. *Let $d(s_0) = s_0 \vee (\neg s_0 \wedge s_1)$ and $d(s_1) = (s_0 \wedge s_1) \vee (\neg s_0 \wedge \neg s_1)$. f_{s_1} has the following table: There are two fixed points of f_{s_1}, namely $(1,1)$ and $(1,0)$. This means*

(x, y)	$(1,1)$	$(1,0)$	$(0,1)$	$(0,0)$
$f_{s_1}(x, y)$	$(1,1)$	$(1,0)$	$(1,0)$	$(0,1)$

that there are two ways of consistently assigning truth values to this hypodox: assigning both s_0 and s_1 as true, or assigning s_0 as true and s_1 as false. Thus, s_0 has to be true in any consistent assignment of truth values to the hypodox.

Therefore, there are hypodoxes for which — even though there are multiple consistent assignments of truth values to all sentences in the hypodox — every one of these assignments gives the same truth value to some particular sentence in the hypodox. This raises a natural question: is this a sufficient reason to believe that this particular sentence should be classified as true or false? Or, put in a more neutral way, could there be a fruitful theory of truth in which such sentences are treated like ordinary (non-paradoxical) truths or falsehoods?

I believe these questions open up rich avenues for further research. They also highlight the robustness of the framework developed in this thesis: by analysing the patterns of fixed points of an ascription function, we find new ways to assess the semantic status of hypodoxical sentences.

5.2 Reduction Operation

In (Rabern, Rabern, and Macauley, 2012), they provide a graph-theoretic sense where the Liar paradox underlies every liar cycle of length n — L_0 = "L_1 is true", L_1 = "L_2 is true", ..., L_{n-1} = "L_0 is not true". However, we also have an intuition that this paradox is not only similar to the Liar paradox, but it just is the Liar paradox after we get rid of the redundant sentences in between. For example, there is really no need for L_1 to be here, because it is just a confirmation of L_2, so we can just substitute L_1 for L_2. Similarly, we can substitute L_2 for L_3, and so on, and finally we can substitute L_{n-1} for $\neg True(L_0)$. In general, if L_i says something about the truth of L_j, then we can substitute that for what L_j says. This can be captured by the tools we have:

Definition 5.8 (0-Reduction)**.** *Let $h(\bar{x}) = (h_0(\bar{x}), h_1(\bar{x}), \ldots, h_n(\bar{x}))$. Then we say $g : \{0,1\}^n \to \{0,1\}^n$ is a 0-reduction of h if $g(\bar{x}) = (g_0(\bar{x}), g_1(\bar{x}), \ldots, g_n(\bar{x}))$, where*

$$g_i(\bar{x}) = h_i(x_0, x_1, \ldots, x_{m-1}, h_m(\bar{x}), x_{m+1}, \ldots, x_n)$$

for some $m \leq n$. That is, we substitute occurrence of x_m in h_i with $h_m(\bar{x})$.

The meaning of the above definition is as follows. Assume s_1 depends on s_2, which is the m-th sentence in the list V_{s_1}. Let $f_{s_1} = (\bar{x}) = (h_0(\bar{x}), h_1(\bar{x}), \ldots, h_n(\bar{x}))$. Then $h_m(\bar{x})$ gives the revised truth value of s_2 according to $\bar{x}$. Then in other coordinates $h_i(\bar{x})$ of $f_{s_1}(\bar{x})$, we can substitute $h_m(\bar{x})$ for x_m. Let us see it work in the Liar circle.

Example 5.9. *Liar circle: let $s_0, s_1, \ldots, s_n \in S$ be such that $d(s_i) = s_{i+1}$ for $i < n$, and $d(s_n) = \neg s_0$. We have $V_{s_0} = \langle s_0, s_1, \ldots, s_n \rangle$ and $f_{s_0}(\bar{x}) = (x_1, x_2, \ldots, x_n, 1 - x_0)$. Then $f_{s_0}(\bar{x}) = (h_0(\bar{x}), h_1(\bar{x}), \ldots, h_n(\bar{x}))$ where $h_i(\bar{x}) = x_{i+1}$ for $i < n$ and $h_n(\bar{x}) = 1 - x_0$. Let us substitute $h_1(\bar{x}) = x_2$ for x_1. Then we have $g^1(\bar{x}) = (x_2, x_2, x_3, \ldots, x_n, 1 - x_0)$. Now we can substitute $g_2^1(\bar{x}) = x_3$ for x_2, and then we have $g^2(\bar{x}) = (x_3, x_3, x_3, \ldots, x_n, 1 - x_0)$. After $n-1$ steps, we have $g^{n-1}(\bar{x}) = (x_n, x_n, x_n, \ldots, x_n, 1 - x_0)$. Finally, we can substitute $g_n^{n-1}(\bar{x}) = 1 - x_0$ for x_n, and we have $g^n(\bar{x}) = (1 - x_0, 1 - x_0, 1 - x_0, \ldots, 1 - x_0, 1 - x_0)$.*

Informally, let us see a Liar circle with 3 sentences: L_0 = "L_1 is true"; L_1 = "L_2 is true"; L_2 = "L_0 is not true". Then we can substitute what L_1 says into L_0 and get L_0' = "L_2 is true". Then we can substitute what L_2 says into L_0' and get L_0'' = "L_0 is not true".

There is a corresponding syntactical operation for this reduction.

Definition 5.10 (Syntactical Reduction)**.** *Let $s_1, s_2 \in S$ be propositional letters in the sentence system (S, d). We say s_2 is a* 0-reduction *of s_1 if $d(s_2)$ can be obtained from $d(s_1)$ by substituting some $t \in V(s_1)$ for $d(t)$.*

Further, we say $s_3 \in S$ is a reduction *of s_1 (or s_1 can be reduced to s_3) if there is a sequence of sentences $s_1, s_2, \ldots, s_\alpha$ such that s_i is a 0-reduction of s_{i-1} for all $i \leq n$.*

Example 5.11. *Take the Liar circle of length 3 where $d(s_0) = s_1, d(s_1) = s_2$, and $d(s_2) = \neg s_0$. Then s_0 can be reduced to s_1 because $d(s_1)$ can be obtained by substituting $d(s_1)$ for s_1 in $d(s_0)$. Similarly, s_1 can be reduced to s_2. Therefore, we also have that s_0 can be reduced to s_2.*

Note that there is still one step from what we want: we want to reduce the above Liar circle to the Liar sentence L_0 = "L_0 is not true", instead of L_0'' = "L_0 is not true". This is reflected in the formal definition because we only defined how to reduce a function $h : \{0,1\}^n \to \{0,1\}^n$ to a function $g : \{0,1\}^n \to \{0,1\}^n$, but to obtain the Liar sentence, we need to reduce a function $h : \{0,1\}^n \to \{0,1\}^n$ to a function $g : \{0,1\} \to \{0,1\}$. From the example, we have already seen a natural way to do this, because $g^n(\bar{x}) = (1 - x_0, 1 - x_0, 1 - x_0, \ldots, 1 - x_0, 1 - x_0)$ actually only depends on the value of x_0. Moreover, the 0-th coordinate of the output is $1 - x_0$, which does not depend on anything but the 0-th coordinate of the input. Therefore, one can forget about all the other coordinates and just take the 0-th coordinate of both the input and the output to obtain the Liar sentence. Thus, there is a way to reduce a function $h : \{0,1\}^n \to \{0,1\}^n$ to a function $g : \{0,1\}^m \to \{0,1\}^m$, where $m < n$.

Definition 5.12 (1-Reduction)**.** *Let $\bar{x} = x_0, \ldots, x_n$ and $\bar{x}' = x_0, \ldots, x_{m-1}, x_{m+1}, \ldots, x_n$, where $0 \leq m \leq n$. Let $h(\bar{x}) = (h_0(\bar{x}), h_1(\bar{x}), \ldots, h_n(\bar{x}))$. Then we say*

$$g(\bar{x}') = (g_0(\bar{x}'), \ldots, g_{m-1}(\bar{x}'), g_{m+1}(\bar{x}'), \ldots, g_n(\bar{x}'))$$

is a 1-reduction *of h if for all $x_0, x_1, \ldots, x_n$, we have $g_i(\bar{x}') = h_i(\bar{x})$ for all $i \neq m$.*

That is, we say $g : \{0,1\}^{n-1} \to \{0,1\}^{n-1}$ is a 1-reduction of $h : \{0,1\}^n \to \{0,1\}^n$ if g is the same as h except that the m-th coordinate of h is removed. In general, for $m \leq n$, we define the set of sentences $g : \{0,1\}^m \to \{0,1\}^m$ which are reductions of $h : \{0,1\}^n \to \{0,1\}^n$ recursively.

Definition 5.13 (Reduction)**.** *Let $m \leq n$ and $h : \{0,1\}^n \to \{0,1\}^n$. We say $g : \{0,1\}^m \to \{0,1\}^m$ is a reduction of h if:*

1. *$m = n$ and $g = h$ or g is a 0-reduction of h; or*
2. *g is a 1-reduction of h;*
3. *there is a sequence of reductions $h_0, h_1, \ldots, h_k$ such that $h_0 = h$ and $h_k = g$.*

Therefore, g is a reduction of h if g can be obtained from h by a sequence of 0-reductions and 1-reductions — i.e., by substituting what a sentence says into another sentence, and when we find a sentence is redundant in the sense that no other sentence depends on it after the substitution, we can remove it.

Now we can formally say that the Liar circle can be reduced to the Liar sentence by showing that the function corresponding to the Liar circle can be reduced to the function corresponding to the Liar sentence.

Lemma 5.14. *Let $h : \{0,1\}^n \to \{0,1\}^n$ be the function corresponding to the Liar circle of length n and $g : \{0,1\} \to \{0,1\}$ be the function corresponding to the Liar sentence — i.e. $h(\bar{x}) = (x_1, x_2, \ldots, x_n, 1 - x_0)$ and $g(x) = 1 - x$. Then g is a reduction of h.*

Proof. We have seen that $h^n(\bar{x}) = (1-x_0, 1-x_0, 1-x_0, \ldots, 1-x_0, 1-x_0)$ can be obtained from h by a sequence of 0-reductions. Notice that h^n only depends on the value of x_0, so we can then perform a sequence of 1-reductions to remove all the other coordinates. □

Similarly, the Truth Teller circle can be reduced to the Truth Teller sentence.

Lemma 5.15. *Let $h : \{0,1\}^n \to \{0,1\}^n$ be the function corresponding to the Truth Teller circle of length n and $g : \{0,1\} \to \{0,1\}$ be the function corresponding to the Truth Teller sentence — i.e. $h(\bar{x}) = (x_1, x_2, \ldots, x_n, x_0)$ and $g(x) = x$. Then g is a reduction of h.*

Then some questions naturally arise:

1. Can we reduce any paradoxical sentences to the Liar and any hypodoxical sentence to the Truth Teller?

2. Clearly the Liar and the Truth Teller cannot be further reduced, but if the previous point is not true, what are all the irreducible sentences?

3. What about paradoxes, like Yablo's, which correspond to an infinitary function?

These are interesting questions worthy of further research.

5.3 Classical Variation of the Theory

Recall from Section 3.4 that if a sentence depends on a paradoxical sentence, our theory will not classify it as true, whereas Kripke's theory sometimes does. For example, let s_0 be the Liar — i.e. $d(s_0) = \neg s_0$. Then strong Kleene version classifies the sentence s_1 with $d(s_1) = s_0 \vee T$ as true is classified as true, since no matter what truth value s_0 takes, one disjunct ($\top$) is already true, and under strong Kleene logic this suffices to make the entire disjunction true. The supervaluation version classifies the sentence s_2 with $d(s_2) = s_0 \vee \neg s_0$ as true, because s_2 holds in any classical model. Although we have argued that caution is warranted when dealing with sentences that depend on a paradox — if these are classified as true, then certain classical principles have to be given up — ultimately, the question is which classical intuition one wishes to preserve. After all, both s_1 and s_2 are classically equivalent to $\top$ — under any classical valuation, they receive

the value 1. One might therefore ask whether the intuition that classically equivalent sentences should be classified as the same could be maintained, and whether our theory can be modified to classify these sentences as true.

In this section, I develop a variant of the proposed theory of truth — henceforth called classical truth — which does just that. It enjoys several desirable features: it is classically consistent, closed under classical equivalence, and preserves semantic status under isomorphism of sentence systems. Moreover, every one of Kripke's three extensions of truth turns out to be a proper subset of the classical truth.

Inspired by the fact that s_1 and s_2 are classically equivalent to $\top$, we may define a sentence s to be classically true if $d(s)$ is classically equivalent to $d(s_0)$ for some true sentence s_0.

In our previous example, s_2 is classically true because we can find an s_4 such that $d(s_4) = \top$. Then $d(s_4)$ is classically equivalent to $d(s_2)$ and s_4 is classified as true; and s_3 is also classically true since it can be reduced to s_2. However, there is a technical issue here: there might not be an s_3 with $d(s_3) = \top$ in some sentence system. This means that we did not give $\top$ a name in the language system we are using. The issue can be easily fixed by extending[5] the original system (S, d) by adding a propositional letter s' that is not in S, and use s' to denote $\top$.

Now, let us formulate classical truth:

Definition 5.16 (Classical Truth)**.** *Let $s \in S$ be a sentence in the sentence system (S, d). We say that s is* **classically true** *if there exists a sentence system (S', d') extending (S, d) and $s_1 \in S'$ such that s_1 is true and $d'(s)$ is classically equivalent to a $d'(s_1)$.*

I believe the above definition is already well motivated, and we will study the formal properties of classical truth. Nevertheless, one might seek to improve it be the following example. Consider s_2 as above, but we add a third sentence s_3 such that $d(s_3) = s_2$. Then $d(s_3)$ is not classically equivalent to any true sentence because it is just a propositional letter. However, we clearly also want to classify s_3 as true if we want to classify s_2 as true, because intuitively s_3 says nothing but "s_2 is true". Using the notion we defined in Section 5.2, we can see that s_3 can be reduced[6] to s_2, which in turn is classically equivalent to $\top$. This leads to the following definition, though we will see shortly that this is problematic.

Definition 5.17 (Reducible to Classical Truth)**.** *Let $s \in S$ be a sentence in the sentence system (S, d). We say that s is* reducible to classical truth *if there exists a sentence system (S', d') extending (S, d) and $s_1 \in S'$ such that s_1 is true and either*

(1) $d'(s)$ is classically equivalent to a $d'(s_1)$; or

[5]Recall Definition 2.37 on the extension of sentence systems.

[6]Recall definition 5.10 on syntactical reduction

(2) there exists $s_2 \in S'$ such that $d'(s_2)$ is classically equivalent to $d'(s_1)$ and s can be reduced to s_2.

Note that reduction reflects reducibility to classical truth, i.e., if a sentence can be reduced to another sentence which is reducible to classical truth, then the first sentence is also reducible to classical truth.

Lemma 5.18. *Let $s_1, s_2 \in S$ be sentences in the sentence system (S, d). Suppose s_1 can be reduced to s_2. If s_2 is reducible to classical truth, then s_1 is also reducible to classical truth.*

Proof. Trivially follows from the definition. □

However, it does not respect classical equivalence because one can find two sentences that are classically equivalent to each other, but one is reducible to classical truth while the other is not.

Lemma 5.19. *There are sentences $s_1, s_2 \in S$ such that s_1 is reducible to classical truth, s_2 is not, and $d(s_1)$ is classically equivalent to $d(s_2)$.*

Proof. Let $d(s_0) = \neg s_0$ be the Liar. Let $d(s_1) = s_0 \vee s_0$ and $d(s_2) = s_0$. Then s_1 is reducible to $\top$, by substituting one of the s_0 by $d(s_0)$, while s_2 could only be reduced to s_0, $\neg s_0$, $\neg\neg s_0, \ldots$, none of which can be equivalent to a true sentence. □

This is very undesirable: while a classically minded people might like to classify an s_2 with $d(s_2) = s_0 \vee \neg s_0$ and an s_3 with $d(s_3) = s_2$ as true, it would be very strange to say that s_1 is true on the ground that one can replace one of the s_0 in $d(s_1)$ by $d(s_0) = \neg s_0$ and keep the other s_0 unchanged.

A way to reconcile this issue is to require that one substitutes *all* instances of a sentence s by $d(s)$ when one performs a reduction:

Definition 5.20 (Uniform Reduction)**.** *Let $s_1, s_2 \in S$ be propositional letters in the sentence system (S, d). We say s_2 is a* uniform 0-reduction *of s_1 if $d(s_2)$ can be obtained from $d(s_1)$ by substituting some $t \in V(s_1)$ for $d(t)$ in all occurrence of t in $d(s_1)$.*

Further, we say $s_3 \in S$ is a uniform reduction *of s_1 (or s_1 can be uniformly reduced to s_3) if there is a sequence of sentences $s_1, s_2, \ldots, s_\alpha$ such that s_i is a 0-reduction of s_{i-1} for all $i \leq n$.*

This resolves the dilemma above, as now $s_0 \vee s_0$ could only be uniformly reduced to $\neg s_0 \vee \neg s_0$, while s_3 can still be uniformly reduced to s_2.

A notion of truth built upon this idea would give:

Definition 5.21 (Uniformly Reducible to Classical Truth)**.** *Let $s \in S$ be a sentence in the sentence system (S, d). We say that s is* **uniformly reducible to classical truth** *if there exists a sentence system (S', d') extending (S, d) and $s_1 \in S'$ such that s_1 is true and either*

(1) $d'(s)$ is classically equivalent to a $d'(s_1)$; or

(2) there exists $s_2 \in S'$ such that $d'(s_2)$ is classically equivalent to $d'(s_1)$ and s can be uniformly reduced to s_2.

Let us now explore the formal properties of this new notion of truth.

We first compare it with other notions of truth proposed in this thesis. Recall that T denotes the set of all truth[7]. Let CT be the set of all classical truth[8] and CT_{UR} the sentences uniformly reducible to classical truth. We have:

Lemma 5.22. *$T \subsetneq CT \subsetneq CT_{UR}$.*

Proof. Follows from the fact that a sentence is classically equivalent to itself. To see that the inclusions are proper, just use the examples given above. □

Next, we show that both CT and CT_{UR} enjoy some desirable properties — both are closed under classical equivalence and are classically consistent.

CT is closed under classical equivalence:

Lemma 5.23. *Let $s_1, s_2 \in S$. If $d(s_1)$ is classically equivalent to $d(s_2)$, and s_1 is classically true, then s_2 is classically true.*

Proof. Trivially follows from the definition of classical truth. □

CT is classically consistent:

Lemma 5.24. *Let $s_1, s_2 \in S$. If $d(s_1)$ is classically equivalent to $\neg d(s_2)$, then s_1 and s_2 cannot both be in CT.*

Proof. Assume for a contradiction that $s_1, s_2 \in CT$. Then there exist sentence systems (S_1, d_1) and (S_2, d_2) extending (S, d) and $s_1' \in S_1$ and $s_2' \in S_2$ such that $d_1(s_1)$ is classically equivalent to $d_1(s_1')$ and $d_2(s_2)$ is classically equivalent to $d_2(s_2')$ and s_1' and s_2' are true. Then there exist valuations $v_1 : S_1 \to \{0, 1\}$ and $v_2 : S_2 \to \{0, 1\}$ induced by the fixed points of $f_{s_1'}$ and $f_{s_2'}$ such that $v_1(d_1(s_1)) = v_1(d_1(s_1')) = 1$ and $v_2(d_2(s_2)) = v_2(d_2(s_2')) = 1$. Since $d_1(s_1) = d(s_1)$ and $d_2(s_2) = d(s_2)$ by the definition of an extension, we have $v_1(d_1(s_1)) = v_1(d(s_1)) = 1$ and $v_2(d_2(s_2)) = v_2(d(s_2)) = 1$. Therefore, $v_2(\neg d(s_2)) = 0 \neq v_1(d(s_1))$. However, then $v_1|_S$ and $v_2|_S$ witness that $d(s_1)$ is not classically equivalent to $\neg d(s_2)$, which is a contradiction. Therefore, s_1 and s_2 cannot both be in CT. □

CT_{UR} is also closed under classical equivalence:

Lemma 5.25. *Let $s_1, s_2 \in S$. If $d(s_1)$ is classically equivalent to $d(s_2)$, and $s_1 \in CT_{UR}$ then $s_2 \in CT_{UR}$.*

[7] Recall Definition 2.26.
[8] Recall definition 5.16

Proof. If s_1 is classically true, then we are done since classical truth is closed under classical equivalence.

Otherwise, let s_1' be a sentence in a sentence system (S', d') extending (S, d) such that $d'(s_1)$ is classically true and s_1 can be uniformly reduced to s_1'. Suppose $t_0, t_1, \ldots, t_\alpha$ is the sequence of uniform 0-reduction of s_1 to s_1', where $t_0 = s_1$ and $t_\alpha = s_1'$. Then we can construct a sequence of uniform 0-reductions $t_0', t_1', \ldots t_\alpha'$ — where $t_0' = s_2$ — by making the same substitution whenever we can and let it remain unchanged otherwise. Formally, assume t_{i+1} is obtained from t_i by substituting $s \in V(t_i)$ for $d(s)$ in all occurrences of s in $d(t_i)$. Then we substitute s for $d(s)$ in all occurrences of s in $d(t_i')$ to obtain t_{i+1}'. Note that if s does not occur in $d(t_i')$, then we have $t_{i+1}' = t_i'$. We prove by induction that $d(t_i')$ is classically equivalent to $d(t_i)$ for all $i \leq \alpha$.

The base case is given by assumption, as $d(t_0') = d(s_2)$ is classically equivalent to $d(t_0) = d(s_1)$.

Assume $d(t_i')$ is classically equivalent to $d(t_i)$ for some $i < \alpha$. Assume also that t_{i+1} is obtained from t_i by substituting $s \in V(t_i)$ for $d(s)$ in all occurrences of s in $d(t_i)$. If $s \in V(t_i')$, then $d(t_{i+1}')$ is obtained from $d(t_i')$ by substituting s for $d(s)$ in all occurrences of s in $d(t_i')$[9]. Let $v : S \to \{0, 1\}$ be any valuation. Let $v' : S \to \{0, 1\}$ be such that it agrees with v on all variable except that $v'(s) = v(d(s))$. Clearly, we have $v(d(t_{i+1})) = v'(d(t_i))$ and $v(d(t_{i+1}')) = v'(d(t_i'))$. Since $d(t_i)$ is classically equivalent to $d(t_i')$, we have $v(d(t_i)) = v'(d(t_i'))$. Therefore, $v(d(t_{i+1})) = v(d(t_{i+1}'))$. This shows that $d(t_{i+1}')$ is classically equivalent to $d(t_{i+1})$ since v is arbitrary.

Let $s_2' = t_\alpha'$. By induction, we have $d(s_2') = d(\alpha')$ is classically equivalent to $d(s_1') = d(\alpha)$. Since s_1' is classically true, we have s_2' is also classically true. Hence, s_2 can also be uniformly reduced to a classical truth. □

CT_{UR} is also classically consistent:

Lemma 5.26. *Let $s_1, s_2 \in S$. If $d(s_1)$ is classically equivalent to $\neg d(s_2)$, then s_1 and s_2 cannot both be in CT_{UR}.*

Proof. If both of them are classically true, then by the previous lemma, they cannot both be in CT. As $CT \subsetneq CT_{UR}$, we have s_1 and s_2 cannot both be in CT_{UR}.

Thus, without loss of generality, assume s_1 is not classically true. Then there exists a sentence system (S', d') extending (S, d) and $s_1' \in S'$ such that s_1' is true and s_1 can be uniformly reduced to s_1'. □

Lastly, let us compare CT and CT_{UR} with Krike's extensions of truth. Recall from Chapter 3 that T_K^W is the extension of the Kripkean truth under weak Kleene logic, T_K^S is the extension of the Kripkean truth under strong Kleene logic, and T_K^{Sv} is the extension of Kripkean truth under supervaluation.

[9]Note that this includes the case where s does not occur in $d(t_i')$

Lemma 5.27. *T_K^S and CT are incomparable.*

Proof. Let s_0, s_1, s_2, s_3 be such that $d(s_0) = \neg s_0$, $d(s_1) = s_0 \vee \neg s_0$ and $d(s_2) = s_0 \vee T$, and $d(s_3) = s_2$. As we have already discussed, s_1 in CT but not in T_K^S. On the other hand, s_3 is in T_K^S but not in CT. This is because s_2 enters the extension of truth under strong Kleene logic, in the first step, and then s_3 will be included in the next step. □

Lemma 5.28. *$T_K^{Sv} \subsetneq CT_{UR}$.*

Proof. We show that $T_K^{Sv} \subseteq CT$ (they cannot be equal as $T \subseteq CT$).

We will show by induction that $X_\alpha \subseteq CT$ and $Y_\alpha \cap CT = \emptyset$ for all α, where (X_α, Y_α) is the extension and anti-extension of truth under the α-th application of the supervaluation Kripke-jump. The basic idea is similar to the proofs in section 3.4, so we will provide a proof sketch.

Firstly, by definition we have:

$$X_{\alpha+1} = \{s \in S \mid [\![d(s)]\!]_v = 1 \text{ for all } v \text{ such that } v(X_\alpha) = \{1\} \text{ and } v(Y_\alpha) = \{0\}\}.$$

Let $s \in X_{\alpha+1}$. If $s \in X_\alpha$ then we are done by induction.

Otherwise, there must be some $t \in X_\alpha$ with $t \in D(s)$. Replace all such t in $d(s)$ with $d(t)$ to obtain a sentence ϕ' which is a uniform reduction. We must have $\phi' \in X_\alpha$. Thus, we are done by induction.

□

Lastly, among the Kripkean theories, it is clear that we have:

Lemma 5.29. *$T_K^W \subsetneq T_K^S \subsetneq T_K^{Sv}$.*

Proof. The inclusion part is a similar induction to above, so we omit. To see that the inclusion is strict, consider $d(s_0) = \neg s_0$, $d(s_1) = s_0 \vee \top$, and $d(s_2) = s_0 \vee \neg s_0$. Then $s_1 \notin T_K^W$ but $s_1 \in T_K^S$. $s_2 \notin T_K^S$ but $s_2 \in T_K^{Sv}$. □

Therefore, collecting the results above and those in Section 3.4, we have this hierarchy of truth extensions:

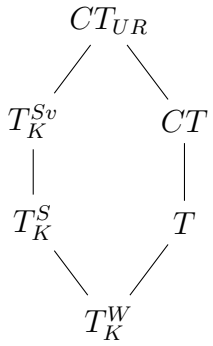


Figure 5.1: Hierarchy of Truth Extensions

Chapter 6

Towards a Theory of Truth in the First Order Language

In this Chapter I briefly outline how the proposed theory of truth can be formulated in the first order language. I will discuss the technical challenges that arise and suggest possible solutions to be further explored in future work.

In (Beringer and Schindler, 2017), they have shown how the notion of a "reference-graph" — developed for propositional languages (Cook 2004, 2014; Rabern, Rabern, and Macauley, 2012) to represent the dependence pattern of sentences — can be extended to first order language. In their case, the presence of quantifiers creates technical complications because it makes the dependence relation much less straightforward than the propositional case.

The same obstacle appears for us to extend our theory of truth to first order language, because an essential ingredient of the proposed theory is also the dependence relation. In the propositional language, we defined[1] that a sentence depends on another sentence if the former contains the latter as a subformula. In a first order predicate language, we cannot define the dependence relation in the same way. For example, consider a formalisation of "all arithmetic sentences are true": $\forall x(Ar(x) \rightarrow True(x))$. Intuitively, we want to say that this sentence depends on all arithmetic sentences. However, the sentence itself does not contain any arithmetic sentence as a subformula. Moreover, neither could the relation be determined by the range of the quantifier, because otherwise the above sentence would depend on all sentences in the language. In fact, there is no syntactic way to define the dependence relation in a first order predicate language, because what a sentence depends on is built on the interpretation of the symbols — as in the above example, Ar — in the language.

To formulate our truth theory in the first order language, we must first supply, for each sentence ϕ a well-defined set D_ϕ of sentences on which ϕ depends; only then can we form its ascription function f_ϕ and analyze its fixed points. A natural candidate for D_ϕ

[1] Recall Definition 2.8.

can be given by employing Leitgeb's (2005) dependence relation, which aims to capture semantic dependence in the first order language. Unfortunately, Leitgeb's definition does not yield a unique dependence set for every sentence, and although he introduces the notion of "essential dependence" to resolve this issue, essential dependence sets fail to exist in some cases.

This chapter proceeds as follows. I begin by showing how our proposed theory can be formulated in the first order language under the assumption that each sentence is already equipped with a well-defined dependence set. From there, I turn to Leitgeb's (2005) dependence relation — and his subsequent notion of essential dependence — and discuss the problem that essential dependence is not defined for all sentences. Lastly, I will suggest two possible ways to resolve the issue, which I believe are worth further exploration in future work.

The language I use is the language of arithmetic $\mathcal{L}$, together with its standard model $\mathbb{N}$. The goal is to extend the language with a truth predicate $True$. The extended language is denoted as $\mathcal{L}^T$. In the extended language, $(\mathbb{N}, \Phi) \models \psi$ means the standard model theoretic satisfaction where Φ is the interpretation of the truth predicate.

6.1 A Theory of Truth in the First Order Language

Let the *dependence set* D_ϕ of ϕ be the set of sentences that ϕ depends on. We assume in this section that there is a well-defined such set for any $\phi \in Sent(\mathcal{L}^T)$.

We now follow the line of Section 2.2 to define the variable set, ascription function, and classify sentences according to the pattern of fixed points of their ascription functions.

Definition 6.1 (Variable Set)**.** *We define the* variable set V_ϕ *for* ϕ *as the set of sentences that* ϕ *depends on plus* ϕ *itself:* $V_\phi = D_\phi \cup \{\phi\}$.

Recall also that we identify V_ϕ with a list of names of sentences ordered by the ordinals, and where $V_\phi(0) = \phi$. Also, we say the index of a sentence $\psi \in V_s$ is the unique ordinal α_ψ such that $V_\phi(\alpha_\psi) = \psi$.

Each sentence ϕ corresponds to a function $f_\phi^{D_\phi} : \{1,0\}^{|V_\phi|} \to \{1,0\}^{|V_\phi|}$ that takes as input a hypothetical truth value of each sentence in V_ϕ and outputs their truth values according to the hypothetical values.

Note that we add a superscript D_ϕ to the ascription function to indicate that it is defined with respect to the dependence set D_ϕ. We do this because it is currently unclear which set we should choose as the dependence set for some sentences.

Definition 6.2 (Ascription Function)**.** *Let* $\phi \in Sent(\mathcal{L}^T)$ *be a sentence in the extended language. We define the* ascription function $f_\phi^{D_\phi} : \{1,0\}^{|D_\phi|} \to \{1,0\}^{|V_\phi|}$ *coordinate-wise as follows:* $f_\phi^{D_\phi}(\bar{x})(i) = 1$ *if* $(\mathbb{N}, \Psi) \models V_\phi(i)$ *where* Ψ *is the set of sentences* ψ *such that* $\bar{x}(\alpha_\psi) = 1$*, and* $f_\phi^{D_\phi}(\bar{x})(i) = 0$ *otherwise.*

That is, given a hypothetical truth assignment $\bar{x}$ to all the sentences in V_ϕ, we take the extension Ψ of the truth predicate to be the set of sentences ψ such that $\bar{x}$ assigns the value 1. Then the output of $f_\phi^{D_\phi}$ is a new truth assignment to the sentences in V_ϕ according to the hypothetical truth assignment: $f_\phi^{D_\phi}(\bar{x})$ assigns the value 1 to the i-th sentence $V_\phi(i) \in V_\phi$ if the standard model $\mathbb{N}$ satisfies the sentence $V_\phi(i)$ under the extension Ψ, and assigns the value 0 otherwise.

Example 6.3. *1. The Liar: let $\phi = \neg True(\ulcorner \phi \urcorner)$ be the Liar sentence. Assume that $D_\phi = \{\phi\}$ — i.e. we have a notion of dependence correctly capturing the fact that the Liar depends on itself. Then $V_\phi = \{\phi\}$ and $f_\phi^{D_\phi} : \{0,1\} \to \{0,1\}$. We have $f_\phi^{D_\phi}(0) = 1$, because $(\mathbb{N}, \emptyset) \models \neg True(\ulcorner \phi \urcorner)$, and $f_\phi^{D_\phi}(1) = 0$, because $(\mathbb{N}, \{\phi\}) \not\models \neg True(\ulcorner \phi \urcorner)$. Thus, we have $f_\phi^{D_\phi}(x) = 1 - x$.*

2. The Truth Teller: let $\phi = True(\ulcorner \phi \urcorner)$ be the Truth Teller sentence. Assume that $D_\phi = \{\phi\}$. Then $V_\phi = \{\phi\}$ and $f_\phi^{D_\phi} : \{0,1\} \to \{0,1\}$. We have $f_\phi^{D_\phi}(0) = 0$, because $(\mathbb{N}, \emptyset) \not\models True(\ulcorner \phi \urcorner)$, and $f_\phi^{D_\phi}(1) = 1$, because $(\mathbb{N}, \{\phi\}) \models True(\ulcorner \phi \urcorner)$. Thus, we have $f_\phi^{D_\phi}(x) = x$.

3. Yablo's paradox: for all $i \in \omega$, let ϕ_i be the formalisation[2] of "for all $j > i$, ϕ_j is not true". Assume that $D_{\phi_0} = \{\phi_i \mid i \in \omega, i > 1\}$. Then $V_{\phi_0} = \{\phi_i \mid i \in \omega\}$ and $f_{\phi_0}^{D_{\phi_0}} : \{0,1\}^\omega \to \{0,1\}^\omega$. One can check that the α-th coordinate of $f_{\phi_0}^{D_{\phi_0}}(\bar{x})$ is given by $f_{\phi_0}^{D_{\phi_0}}(\bar{x})(\alpha) = min\{1 - x_i \mid \alpha < i\}$.

Note that these functions are exactly the same as ascription functions of the corresponding sentences in the propositional language[3].

The classification is also the same as in the propositional case.

Definition 6.4 (Classisfication of Ascription)**.** *Let $\phi \in Sent(\mathcal{L}^T)$ be a sentence in the extended language. We say that the ascription $f_\phi^{D_\phi}$ is:*

1. **successful** *if the equation $\bar{x} = f_\phi(\bar{x})$ has a unique solution;*
2. **paradoxical** *if the equation $\bar{x} = f_\phi(\bar{x})$ has no solution;*
3. **hypodoxical** *if the equation $\bar{x} = f_\phi(\bar{x})$ has multiple solutions;*

We proceed to define the notion of a hereditarily successful ascription.

Definition 6.5 (Agreement)**.** *Let $\phi \in Sent(\mathcal{L}^T)$ be a sentence in the extended language and let $\psi \in V_\phi$, where $f_\phi^{D_\phi}$ and $f_\psi^{D_\psi}$ are both successful. We say that the ascription function $f_\phi^{D_\phi}$ agrees with the ascription function $f_\psi^{D_\psi}$ if the solution of the equation $\bar{x} = f_\phi^{D_\phi}(\bar{x})$ agrees with the solution of the equation $\bar{y} = f_\psi^{D_\psi}(\bar{y})$ in the sense that $x_{\alpha_\gamma} = y_{\alpha_\gamma}$ for all $\gamma \in V_t$.*

[2]There are at least two ways this can be done: either extending the language (Leitgeb, 2005: p. 164), or using a generalization of the diagonal lemma(Beringer and Schindler, 2016: pp. 3 - 4).

[3]Recall Example 2.13

Definition 6.6 (Hereditarily Successful Ascription)**.** *Let $\phi \in Sent(\mathcal{L}^T)$ be a sentence in the extended language. We say that the ascription $f_\phi^{D_\phi}$ is* **hereditarily successful** *if it is:*

1. *successful;*
2. *for any $\psi \in D_\phi$, $f_\psi^{D_\psi}$ is successful; and*
3. *$f_\phi^{D_\phi}$ agrees with $f_\phi^{D_\psi}$ for any $\psi \in V_\phi$.*

Again, we have this simplified definition of hereditarily successful ascription function (the proof is completely similar so we omit):

Lemma 6.7. *Let ϕ be a sentence in the extended language. The ascription $f_\phi^{D_\phi}$ is hereditarily successful if and only if it is successful and for any $\psi \in V_\phi$, $f_\psi^{D_\psi}$ is successful.*

Finally, we have the classification of sentences in the extended language $\mathcal{L}^T$.

Definition 6.8. *Let ϕ be a sentence in the extended language. We say that ϕ is:*

1. **paradoxical** *if f_ϕ is paradoxical;*
2. **hypodoxical** *if f_ϕ is hypodoxical;*
3. **true** *if f_ϕ is successful and the solution of the equation $\bar{x} = f_\phi(\bar{x})$ satisfies $\bar{x}(0) = 1$.*
4. **false** *if f_ϕ is successful and the solution of the equation $\bar{x} = f_\phi(\bar{x})$ satisfies $\bar{x}(0) = 0$;*

Example 6.9. *Making the same assumptions in Example 6.3, we have that the Liar and Yablo's paradox is paradoxical, while the Truth Teller is hypodoxical.*

Let us close this section by reflecting on our key assumption: each sentence ϕ comes with a single, well-defined dependence set D_ϕ. While the uniqueness of D_ϕ is certainly desirable, we can weaken this requirement without affecting our classification. This is because what truly matters is the fixed-point behaviour of the resulting ascription functions. Suppose, for example, ϕ has a collection of candidate dependence sets $\{D_\phi^i \mid i \in \omega\}$. It suffices to require that for any i, j the functions $f_\phi^{D_\phi^i}$ and $f_\phi^{D_\phi^j}$ share the same fixed-point pattern — they must be both successful, paradoxical, hypodoxical, or hereditarily successful[4]. Whether this weaker, "equivalent up to fixed-point pattern" condition can be met remains open, but it may reduce some burden for future work.

[4]When they are both hereditarily successful, we need to check that they assign the same value to ϕ.

6.2 Leitgeb's Dependence Relation

In this section, I discuss the dependence relations proposed by Leitgeb's (2005). We will see that the notion of "essential dependence" provides a promising candidate for the dependence set to be used in our theory of truth. In fact, for a large variety of sentences (including the Liar, the Truth Teller, and Yablo's paradox), this notion yields a unique dependence set, which can be used to generate the ascription function and to correctly classify those sentences. Nevertheless, there are sentences that do not essentially depend on any set of sentences. We discuss this issue and propose two possible ways to resolve it in the next section.

I will follow the notations in Meadows (2012), and I omit proofs of results that already appeared in Leitgeb (2005).

Definition 6.10 (Leitgeb's Dependence)**.** *A sentence* ϕ depends on *a set of sentences* Φ *if for all* Ψ*, we have* $(\mathbb{N}, \Psi) \models \phi$ *if and only if* $(\mathbb{N}, \Phi \cap \Psi) \models \phi$*.*

To gain some intuition for this concept, let us call a sentence true if it is included in the extension of $True$ and false otherwise. According to this definition, the literal meaning of "ϕ depends on Φ" is this: if ϕ holds (or does not hold) under the assumption that all sentences in Φ are true, then ϕ still holds (or does not hold) if we alter the truth value of all sentences in $\Psi \setminus \Phi$ to false (and the reverse direction also holds). In other words, whenever two candidate extensions of $True$ agree on the truth-values of all sentences in Φ, then whether ϕ holds remains the same even if we force every sentence outside Φ to be false.

There appears to be an asymmetry in the definition. In order to keep the truth of ϕ unchanged, it is of course natural to ask that any candidate extensions of $True$ agree on the truth of all sentences that ϕ depends on, but why should we put all other sentences to false, instead of — for example — putting them to true? If other sentences are irrelevant, it should not matter what we do with them. Thus, there seems to be a stronger notion of dependence:

Definition 6.11. *A sentence* ϕ strongly depends *on a set of sentences* Φ *if for all* Ψ*, we have* $(\mathbb{N}, \Phi) \models \phi$ *if and only if for any* $\Delta \subseteq \Phi^c$*,* $(\mathbb{N}, (\Phi \cap \Psi) \cup \Delta) \models \phi$*.*

That is, we should feel free to put into the extension of truth any sentence that is not relevant, and the truth value of ϕ still should not change.

However, this is equivalent to the original definition, because the arbitrariness of Ψ in the original definition already allows us to put any irrelevant sentence into the extension.

Lemma 6.12. *For any sentence* ϕ*, we have* ϕ *depends on* Φ *if and only if* ϕ *strongly depends on* Φ*.*

Proof. $\Rightarrow$: Suppose ϕ depends on Φ. Let Ψ be any set of sentences. We need to show $(\mathbb{N}, \Psi) \models \phi$ if and only if for any $\Delta \subseteq \Phi^c$, $(\mathbb{N}, (\Phi \cap \Psi) \cup \Delta) \models \phi$.

1. Suppose $(\mathbb{N}, \Psi) \models \phi$ and let $\Delta \subseteq \Phi^c$. Since ϕ depends on Φ, we have $(\mathbb{N}, \Phi \cap \Psi) \models \phi$. As $\Delta \subseteq \Phi^c$, we have $\Phi \cap \Psi = ((\Phi \cap \Psi) \cup \Delta) \cap \Psi$. Therefore, $(\mathbb{N}, ((\Phi \cap \Psi) \cup \Delta) \cap \Psi) \models \phi$. By the order direction of the assumption that ϕ depends on Φ, we have $(\mathbb{N}, (\Phi \cap \Psi) \cup \Delta) \models \phi$.

2. Suppose $(\mathbb{N}, (\Phi \cap \Psi) \cup \Delta) \models \phi$ and let $\Delta \subseteq \Phi^c$. Since ϕ depends on Φ, we have $(\mathbb{N}, ((\Phi \cap \Psi) \cup \Delta) \cap \Phi) \models \phi$. By the same calculation as above, we have $(\mathbb{N}, \Phi \cap \Psi) \models \phi$. Then using the other direction of the assumption that ϕ depends on Φ, we have $(\mathbb{N}, \Psi) \models \phi$.

$\Leftarrow$: This is trivial — just take $\Delta = \emptyset$. □

Nevertheless, the issue with this definition is that a sentence can depend on several distinct sets of sentences. Even worse (for our purpose of using it to generate the ascription function), according to Leitgeb's definition, every sentence depends on the set of all sentences in the language. In fact, as pointed out by Leitgeb:

Lemma 6.13. *For any sentence ϕ, the collection of all sentences it depends on is a filter. That is,*

1. *If ϕ depends on Φ, and $\Phi \subseteq \Psi$, then ϕ depends on Ψ.*

2. *If ϕ depends on Φ and ϕ depends on Ψ, then ϕ depends on $\Phi \cap \Psi$.*

By clause (1), we can always add arbitrarily many irrelevant sentences to the set of sentences that a sentence depends on. Leitgeb (2005) realises this issue and introduces the notion of essential dependence:

Definition 6.14 (Essential Dependence)**.** *ϕ depends on Φ* essentially *if ϕ depends on Φ and for any Ψ, if ϕ depends on Ψ then $\Phi \subseteq \Psi$.*

That is, an essential dependence set of ϕ is the smallest set of sentences on which ϕ depends.

This set does not always exist[5], but if it does, it is unique.

Lemma 6.15. *If ϕ depends on Φ essentially, then Φ is the unique such set.*

The essential dependence set thus offers a natural candidate for D_ϕ in our theory. Its only departure from the theory we have developed in the propositional language is that

[5]I will discuss this issue in more detail in the next section. For now, it suffices to keep in mind that some sentences do not have an essential dependence set.

the essential dependence relation is not transitive[6] — recall that in order to define the ascription function for ϕ, we need to collect all sentences relevant to ϕ, including those to which it refers only indirectly. However, just as in Section 2.2, we can recover transitivity by replacing each essential dependence set with its transitive closure under the essential dependence relation.

By slightly abusing language, if ϕ depends on Φ essentially, and $\psi \in \Phi$, we will also say that ϕ depends on ψ essentially.

Definition 6.16. *We first define a relation $R \subseteq Sent(\mathcal{L}^T) \times Sent(\mathcal{L}^T)$ on the set of sentences as follows: $\phi R \psi$ if and only if ϕ depends on ψ essentially.*

Let R^* be the transitive closure of R_ϕ.

Definition 6.17. *We define the dependence set for ϕ as $D_\phi = \{\psi \in Sent(\mathcal{L}^T) \mid \phi R^* \psi\}$.*

Remark 6.18. *If some sentence ϕ does not have an essential dependence set, then according to our definition $D_\phi = \emptyset$. Then by definition 6.2, $f_\phi^{D_\phi}$ is also the empty function. In this case, for now, we simply say that ϕ does not have an ascription function, and we do not classify this sentence.*

This yields a transitive dependence set suitable for defining ascription functions and studying fixed point behaviour in our theory of truth as outlined in the previous section.

Moreover, note that this definition does give us what we want in the important cases of the Liar, the Truth Teller, and Yablo's paradox.

Example 6.19.

1. *The Liar: let $\phi = \neg True(\ulcorner \phi \urcorner)$ be the Liar sentence. The essential dependence set of ϕ is just $\{\phi\}$. Thus, we have $D_\phi = \{\phi\}$.*

2. *The Truth Teller: let $\phi = True(\ulcorner \phi \urcorner)$ be the Truth Teller sentence. The essential dependence set of ϕ is also just $\{\phi\}$. Then we have $D_\phi = \{\phi\}$.*

3. *Yablo's paradox: for all $i \in \omega$, let ϕ_i be the formalisation of "for all $j > i$, ϕ_j is not true". Then the essential dependence set of ϕ_i is $\{\phi_j \mid j > i\}$. Thus, we have $D_{\phi_0} = \{\phi_i \mid i \in \omega, i > 1\}$.*

These are exactly what we need in Example 6.3. Hence, the truth theory developed in the previous section applies to these cases and produces the correct classification of those sentences.

Therefore, despite the limitation that the theory applies only to sentences with an essential dependence set, it already covers a wide range of cases — including many of the most important examples in the literature.

[6] See Example 12 in (Leitgeb, 2005: p. 164) for a concrete counterexample.

6.3 Sentences Without Essential Dependence Sets

As we have already mentioned, there are sentences which do not essentially depend on any set of sentences. In this section, we present one such example and two potential ways to resolve the issue.

Consider the nested Yablo's paradox given in (Beringer and Schindler, 2016, 2017), where ϕ_n is the formalisation of "there is an $m > n$ such that for all $k > m$, ϕ_k is not true".

Let us first see informally that this is indeed a paradox. Suppose that we could assign each sentence in the nested Yablo sequence a classical truth value in line with T-schema. If we assign every one of them false, then the first sentence ϕ_0 — which asserts that "there is some point beyond which all sentences are false" — would turn out to be true, contradicting our assignment. Therefore, at least one sentence must be true. Assume ϕ_n is one such sentence. By what that sentence says, there must then be some later point $m > n$ such that every sentence after m is false. In particular ϕ_{m+1} must be false. However, ϕ_{m+1} also claims that beyond some point all later sentences are false, and since we have set all those later sentences to false ϕ_{m+1} must itself be true. We reach a contradiction again. Hence, it is impossible to assign truth values to all ϕ_n's consistently, and thus the nested Yablo's paradox is indeed a paradox.

We now illustrate that ϕ_0 does not have an essential dependence set. First, as observed by (Beringer and Schindler, 2017), ϕ_0 depends on all sets of the form $\{\phi_k \mid k \geq m\}$ for any $m \geq 1$. Thus, ϕ_n does not essentially depend on any set of sentences because the empty set is the only set contained in all the above sets, while ϕ_n does not depend on the empty set.

The first way to potentially solve this problem is to use a notion of "irrelevance".

Let us first reflect on whether Leitgeb's dependence relation captures our intuitive idea of dependence in the nested Yablo's paradox. We have seen that ϕ_0 depends on $\{\phi_k \mid k \geq m\}$ for all $m \geq 1$. Nonetheless, this is in some sense intuitively acceptable — given the nested structure of the sentences, it refers to all sentences ϕ_k for $k \geq 1$ because of the existential quantifier, but any finite initial segment $\{\phi_k \mid k < m\}$ of the sequence have no influence on the truth of ϕ_0 because of the universal quantifier. Thus, one may freely include — or omit — any such initial segment from the dependence set of ϕ_0 without worrying about whether the resulting set correctly captures the dependence relation.

However, there is a greater problem: by Lemma 6.13, ϕ_0 also depends on $\{\phi_k \mid k \geq 1\} \cup \{1 = 1\}$ and $\{\phi_k \mid k \geq 1\} \cup \{\phi\}$ where ϕ is the liar. This is clearly not acceptable as one should at least be able to exclude the completely irrelevant sentences from the dependence set of ϕ_0. However, unable to use essential dependence in this case, we cannot tell the difference between these two sets and the sets $\{Y_k \mid k \geq m\}$ for all $m \geq 1$. They are all dependence sets of Y_0, and we are not able to distinguish them by set inclusion.

We aim for a notion of dependence that can help restrict the dependence set of Y_0, so that only sets of the form $\{Y_k \mid k \geq m\}$ are included.

Definition 6.20. *A sentence ϕ is irrelevant to ψ is there is Φ such that ϕ depends on Φ, Ψ such that ψ depends on Ψ, and $\Phi \cap \Psi = \emptyset$.*

That is, if two sentences have some disjoint dependence sets, then they are irrelevant to each other.

Example 6.21. *Clearly, in the nested Yablo's paradox, the sentence $1 = 1$ and the Liar are irrelevant to ϕ_0.*

We want to exclude the irrelevant sentences from the set of sentences that a sentence depends on.

Definition 6.22. *ϕ* depends* *on Φ if ϕ depends on Φ and for any $\psi \in \Phi$, ψ is not irrelevant to ϕ.*

That is, if ϕ depends* on Φ, then Φ cannot contain any sentence irrelevant to ϕ.

Example 6.23. *In the nested Yablo's paradox, ϕ_0 depends* on $\{\phi_k \mid k \geq 1\}$. In fact, ϕ_0 depends* on Φ for any Φ containing a tail[7] of Y_k's.*

It would be great if these are all the sets that Y_n depends* on. However, unfortunately, this is not the case. Consider a sentence ψ which says "there is an $m > 0$ such that for all $k > m$, Y_k is not true, and the Liar is true". That is, it says what ϕ_0 says but claims additionally that the Liar is true. Then this sentence is relevant to ϕ_0, because for any Φ such that ϕ_0 depends on and any Ψ such that ψ depends on Ψ, there is $m \in \omega$ large enough such that $\phi_m \in \Phi \cap \Psi$.

Thus, ϕ_0 depends* on $\{\phi_k \mid k \geq 1\} \cup \{\psi\}$. One can see that this new notion faces almost the same difficulty as the original one — although the irrelevant sentences like the Liar itself cannot be added into the dependence set, we can always first add it as a conjunct to a relevant sentence, and then add the entire conjunction to the dependence set.

Therefore, we need a finer-grained notion of irrelevance to avoid this issue. Whether pursuing ever more refined notions of irrelevance would eventually solve the question or lead to an infinite regress is a question for future work.

The second way to potentially solve the problem is to apply some metatheoretical principle — to decide the dependence set of a sentence with nested quantifiers, only the outermost quantifier matters.

I have argued that it is intuitively acceptable to say ϕ_0 depends on $\{\phi_k \mid k \geq 1\}$, because ϕ_0 does refer to all of them via the outermost existential quantifier. Let us further

[7]That is, there is some $n \in \omega$ such that $\{\phi_k \mid k \geq n\} \subseteq \Phi$.

see that taking this set as the dependence set of ϕ_0 yields an ascription function and hence classification of ϕ_0 that gives the correct reasoning.

Assume $D_{\phi_0} = \{\phi_k \mid k \geq 1\}$. Then $V_{\phi_0} = \{\phi_k \mid k \geq 0\}$ and $f_{\phi_0}^{D_{\phi_0}} : \{0,1\}^\omega \to \{0,1\}^\omega$ is such that

$$f_{\phi_0}^{D_{\phi_0}}(\bar{x})(n) = \begin{cases} 1, & \text{if there is } m > n \text{ such that for all } k > m, \bar{x}(k) = 0; \\ 0, & \text{otherwise.} \end{cases}$$

Let us show that this function does not have any fixed point and observe that the proof parallels how we would decide the nested Yablo's paradox is paradoxical informally. Firstly, $\bar{0} \in \{0,1\}^\omega$ is not a fixed point, because $f_{\phi_0}^{D_{\phi_0}}(\bar{0})(0) = 1$. Therefore, assume $\bar{x}$ is a fixed point of $f_{\phi_0}^{D_{\phi_0}}$ — i.e. $\bar{x} = f_{\phi_0}^{D_{\phi_0}}(\bar{x})$. Then $\bar{x}(n) = 1$ for some $n \in \omega$ and hence $f_{\phi_0}^{D_{\phi_0}}(\bar{x})(n) = 1$. By definition of $f_{\phi_0}^{D_{\phi_0}}$, there must be some $m > n$ such that we must have $\bar{x}(k) = 0$ for all $k > m$. However, then $\bar{x}(m+1) = f_{\phi_0}^{D_{\phi_0}}(\bar{x})(m+1) = 1$, contradiction. This argument matches exactly with the informal reasoning we gave at the beginning of this section.

Therefore, for both intuitive and pragmatic reasons, it is acceptable to ignore the inner quantifiers in a sentence with nested quantifiers like "there is an $m > 0$ such that for all $k > m$, ϕ_k is not true" and to say that it depends on the same set of sentences as the sentence "there is an $m > 0$ such that ϕ_m is not true" depends.

Furthermore, as observed in (Leitgeb, 2005: p. 164), the latter sentence does have an essential dependence set, $\{\phi_k \mid k \geq 1\}$ — which is exactly what we need.

This suggests that we can use a two-step procedure to resolve the problem with nested quantifiers: first, we appeal to some metatheoretical principle to transform the sentence into one with only one quantifier — for example, by ignoring the inner quantifiers; then we apply Leitgeb's essential dependence relation to the resulting sentence to obtain a unique dependence set.

There remain two critical questions for this approach:

1. How to formulate the metatheoretical principle for sentences with nested quantifiers in general?

2. Is it true that all sentences having only one quantifier have an essential dependence set? In (Leitgeb, 2005; Beringer and Schindler, 2016, 2017), only examples with nested quantifiers are shown to have no essential dependence set, and it remains open to characterise the class of sentences that have an essential dependence set.

These questions will be answered in future work.

Therefore, although there are sentences having no essential dependence set so that the theory of truth developed in the previous sections does not apply, we have two potential approaches to resolve this issue. The first way is to use a notion of irrelevance to exclude

irrelevant sentences from the dependence set of a sentence. The second way is to appeal to some metatheoretical principle to transform the sentence into one that has an essential dependence set.

These potential solutions lead to further research questions interesting in their own right — whether we can formally characterise when two sentences are irrelevant to each other; and how to characterise sentences which do have an essential dependence set. Furthermore, the proposed theory can already deal with important examples like the Liar, the Truth Teller, and Yablo's paradox. Therefore, the proposed theory of truth has robust potential for future research.

Chapter 7

Conclusion

In this thesis, I have introduced a theory of truth grounded in an intuitive procedure for classifying the semantic status of a sentence — for any sentence, we identify sentences relevant to it and ask whether there is a unique way of consistently assigning truth values to all of them. This procedure is modelled by corresponding each sentence with a variable set and an ascription function. The classification is then given by the pattern of fixed points of the ascription function.

I have formally developed the theory of truth in an infinite propositional language, and we see that it has certain advantages over the two dominant theories of truth in the literature: Kripke's theory of truth and the Revision theory of truth. Moreover, the theory has robust applications in the study of semantic paradox.

Finally, I sketched how the framework can be extended to the first order language by using the notion of essential dependence relation proposed by Leitgeb (2005). While some sentences do not have an essential dependence set, many of the central examples in the literature — Liar, Truth Teller, Yablo — do admit one and are correctly handled by the theory. Furthermore, I proposed two strategies for filling the remaining gaps, each leading to further research questions interesting in their own right.

In the future, I will further pursue the research questions raised in this thesis, including a more detailed comparison with Kripke's theory, the application of the theory to the study of hypodoxes, and the completion of the first-order extension.

Bibliography

1. Beringer, T. and Schindler, T. (2016). Reference Graphs and Semantic Paradox.

2. Beringer, T. and Schindler, T. (2017). A Graph-Theoretic Analysis of the Semantic Paradoxes. The Bulletin of Symbolic Logic, 23(4), pp. 442-492. doi:10.1017/bsl.2017.37.

3. Cook, R.T. (2004). Patterns of Paradox. Journal of Symbolic Logic, 69(3), pp.767-774. doi:https://doi.org/10.2178/jsl/1096901765.

4. Cook, R.T. (2011). The No-No Paradox Is a Paradox. Australasian Journal of Philosophy, 89(3), pp.467-482. doi:https://doi.org/10.1080/00048402.2010.500671.

5. Cook, R.T. (2014). The Yablo Paradox. OUP Oxford.

6. CSCI E-80 (2021) Knights. Available at: https://cs50.harvard.edu/extension/ai/2021/fall/projects/ (Accessed: 09 June 2025).

7. Halbach, V. (2014). Axiomatic Theories of Truth. Cambridge University Press.

8. Halbach V, Leigh GE. (2024) The Road to Paradox: A Guide to Syntax, Truth and Modality. Cambridge University Press.

9. Field, H. (2008). *Saving Truth From Paradox*. OUP Oxford.

10. Gupta A. and Belnap, N.D. (1993). *The Revision Theory of Truth*. Cambridge, Mass.: MIT Press.

11. Herzberger, H.G. (1982a). Naive Semantics and the Liar Paradox. *The Journal of Philosophy*, 79(9), p.479. doi:https://doi.org/10.2307/2026380.

12. Herzberger, H.G. (1982b). Notes on naive semantics. Journal of Philosophical Logic, 11(1), pp.61-102. doi:https://doi.org/10.1007/bf00302339.

13. Kreisel, G., (1965). Mathematical Logic. In T. L. Saaty (ed.), Lectures on Modern Mathematics, vol. III, pp. 95–195. New York: John Wiley.

14. Kripke, S. (1975). Outline of a Theory of Truth. The Journal of Philosophy, 72(19), p.690. doi:https://doi.org/10.2307/2024634.

15. Leitgeb, H. (2005). What Truth Depends on. Journal of Philosophical Logic, 34(2), pp.155-192. doi:https://doi.org/10.1007/s10992-004-3758-3.

16. Meadows, T. (2012). Truth, Dependence and Supervaluation: Living with the Ghost. Journal of Philosophical Logic, 42(2), pp.221-240. doi:https://doi.org/10.1007/s10992-011-9219-x.

17. Priest G. (2008). An Introduction to Non-Classical Logic: From If to Is. 2nd ed. Cambridge University Press.

18. Rabern, L., Rabern, B. and Macauley, M. (2012). Dangerous Reference Graphs and Semantic Paradoxes. Journal of Philosophical Logic, 42(5), pp.727-765. doi:https://doi.org/10.1007/s10992-012-9246-2.

19. Smith, P. (2013). An Introduction to Gödel's Theorems. Cambridge University Press.

20. Smullyan, R.M. (1978). What is the Name of this Book? Prentice Hall.

21. Sorensen, R. (2001). Vagueness and Contradiction. Clarendon Press.

22. Tarski, A. (1956). The Concept of Truth in Formalized languages. In Logic, semantics, metamathematics. Oxford,: Clarendon Press. pp. 152–278.

23. Tent K, Ziegler M (2012). A Course in Model Theory. Cambridge University Press.

24. Yablo, S. (1993). Paradox without Self-Reference. Analysis, 53(4), pp.251-252. doi:https://doi.org/10.1093/analys/53.4.251.