

How well does Hansen's theory model truth and avoid paradox?

A reply to Casper S. Hansen, "Grounded Ungroundedness"

Marleen Westerik

August 29, 2014

Supervisor:
prof. dr. Albert Visser

Credits:
15 EC Bachelor thesis

Contents

1	Introduction	4
1.1	The Liar Paradox	4
1.1.1	The Simple Liar	5
1.1.2	The Strengthened Liar	5
1.1.3	Yablo’s Paradox	6
1.1.4	The Empirical Liar	6
1.2	Solutions to the Liar	7
1.2.1	Towards Hansen (2014)	7
1.2.2	Other approaches	9
1.3	Research question	11
1.4	Relevance for Artificial Intelligence	11
1.4.1	Paradoxes	11
1.4.2	Humans	12
1.4.3	Truth	12
1.5	Structure	13
2	Hansen’s theory	14
2.1	Flaws in Kripke’s theory	14
2.1.1	Undefinedness	14
2.1.2	Connectives	15
2.1.3	Names	16
2.2	Hansen’s theory	16
2.2.1	Intuitive account	16
2.2.2	Formal theory	17
2.3	Clarifications	19
2.3.1	Quoted in	19
2.3.2	α -equivalence	20
2.4	Undefined sentences	20
2.4.1	Undefined at higher levels	20
2.4.2	The Disjunctive Liar	21
2.4.3	Sentences containing quotes	21
2.4.4	Undefined subformulas	22
2.4.5	Satisfying E3	22
2.4.6	Undefined formulas and subformulas	22

2.4.7	Revenge	23
2.4.8	Self-reference	23
2.4.9	Yablo's paradox	23
2.5	E2	24
2.6	Conclusion	26
3	Hansen and Kripke	27
3.1	Kripke's grounded sentences	27
3.1.1	Technical requirements	28
3.1.2	Sentences do not become undefined too early	28
3.1.3	K-grounded sentences and proper truth values	29
3.2	Kripke's other fixed points	32
3.3	Semantic equivalences in Hansen and Kripke	33
3.3.1	Propositional equivalences	34
3.3.2	Quantified equivalences	35
3.3.3	Implication	36
3.3.4	Truth	37
3.4	Conclusion	37
4	A critical view	39
4.1	Undefined formulas	39
4.2	Semantic equivalences	40
4.3	Choice of negation	41
4.4	Strong Kleene	42
4.5	Stability of truth values	44
4.6	The Leibniz law	45
4.7	Domain constants	46
4.8	Satisfaction	47
4.8.1	The satisfaction predicate	47
4.8.2	Hansen-satisfaction	47
4.8.3	Names	50
4.8.4	Compositionality	51
4.9	E4	51
4.9.1	E4 and circularity	52
4.9.2	E4 and infinity	52
4.9.3	A new E4	53
4.10	Intuitively true generalisations	55
4.11	Conclusion	56
5	Conclusion	58
5.1	Summary	58
5.1.1	Hansen's theory	58
5.1.2	Hansen and Kripke	59
5.1.3	A critical view	59
5.2	About this thesis	60
5.2.1	Method	60

5.2.2	Objective	60
5.3	Further research	60
5.3.1	Philosophy	60
5.3.2	Effects of adaptations	61
5.3.3	Applications	61
5.3.4	Satisfaction	61
5.4	Impact on Artificial Intelligence	61
5.5	How well does Hansen's theory model truth and avoid paradox? .	62

Chapter 1

Introduction

“You say you are lying but if everything you say is a lie then you are telling the truth but you cannot tell the truth because everything you say is a lie but you tell the truth but you cannot for you lie. Illogical! Illogical!” - Android Norman shortly before he shuts down (Star Trek, episode 37, 1968)

If we ask which things are true, we need to know first what it means for something to be true, and the fact of the matter is: we don't know. We can use formal logic to make all sorts of claims about the world, and given a model of the world, those claims can be evaluated to receive a truth value in the meta-language. The concept of truth itself however remains a particularly challenging concept to define in the object-language: claiming that it is true that there are nine million bicycles in Beijing turns out to be much more ambitious than one may initially think. A number of solutions have been suggested over the years, among which the one found in Hansen (2014), which will be the subject of this thesis.

1.1 The Liar Paradox

The reason for the difficulty in defining truth can be found in a family of paradoxes referred to as the Liar Paradox. These paradoxes have in common that they all use the concept of truth in such a way that it seems to become impossible to say anything sensible about their truth value at all. If one wishes to distinguish between the different members of the family, the name Liar Paradox is often used to refer to the Simple Liar, and several other types of liar paradoxes are also known by their own name. This section briefly introduces a few of the most influential liar paradoxes.

1.1.1 The Simple Liar

The simple liar says of itself that it is false:

The Liar: The Liar is false

It is paradoxical because if it were true, by virtue of its meaning, it would have to be false. But if it were false, by virtue of its meaning, it would have to be true. The Liar is true if and only if it is false. The consequences of this paradoxality are discussed in more depth in the next paragraph.

1.1.2 The Strengthened Liar

The Strengthened Liar is a version of the liar that uses ‘not true’ instead of false. That is:

The Strengthened Liar: The Strengthened Liar is not true.

Again it follows that if the sentence is true, then it cannot be true, and if it is not true, it must be true. Also, by using ‘not true’ instead of ‘false’, it includes any other truth values besides true and false one may want to introduce.¹

The (Strengthened) Liar leads to absurdity: contradiction without assumptions. The following natural deduction style proof shows the structure of the reasoning, where the Strengthened Liar is formalised as $\neg T(c_l)$, with T the truth predicate and c_l a constant such that $I(c_l) = \neg T(c_l)$. This formalisation of the Strengthened Liar will be discussed in more detail later.

$$\frac{\frac{\frac{\frac{\frac{[T(c_l)]^1}{T(\neg T(c_l))} R}{\neg T(c_l)} \neg E}{\perp} \perp}{T(c_l) \vee \neg T(c_l)} LEM}{\perp} \perp}{\frac{\frac{\frac{[\neg T(c_l)]^1}{\neg T(\neg T(c_l))} SI}{\neg \neg T(c_l)} R}{\perp} \perp}{[\neg T(c_l)]^1} \neg E}{\perp} \perp} \vee E^1$$

In this proof, formulas between quotes are names for those formulas. *SI* stands for ‘Substitutivity of Identity’, that is: different names of the same formula can be substituted for each other. *R* stands for ‘Release’, the idea that if you can claim that a certain formula is true, then you can claim the formula itself also. *LEM* stands for ‘Law of Excluded Middle’² The proof shows that seemingly natural rules lead to absurdity. Devising *EF SQ* (‘Ex Falso Sequitur Quodlibet’), one can now infer any formula ϕ :

$$\frac{\perp}{\phi} \text{ EF SQ}$$

¹As will be discussed in section 2.1.1, the Strengthened Liar forms a persistent problem for so-called paracomplete theories of truth. Hansen’s theory is paracomplete, but does handle the Strengthened Liar in an elegant fashion.

²This proof is intended to represent the intuitive Liar reasoning. A slightly less intuitive proof that does not make use of *LEM* is also possible and can be found in Visser (1989), p162. This shows that denying the validity of *LEM* does not block the Liar.

This shows that if you just add a truth predicate to a language with some rather intuitive rules to govern it, you get trivialisation of the system: *any* formula can be derived. This means that at least one of the rules used to achieve this result must be abandoned.

1.1.3 Yablo's Paradox

Most versions of the Liar employ self-reference, which may suggest that self-reference is necessarily the cause for Liar-type paradoxes. Self-reference does not necessarily lead to paradox, for instance the sentence "This sentence is in English" is perfectly unproblematic. The problem lies with sentences whose truth value needs to be established before it can be established. Yablo (1993) showed that paradox can also be created without self-reference. He came up with a paradox that has since been known as Yablo's Paradox which does not employ self-reference:

- (S_1) : For all $k > 1$, S_k is untrue
 (S_2) : For all $k > 2$, S_k is untrue
 (S_3) : For all $k > 3$, S_k is untrue
 \vdots

If a sentence S_i in this infinite sequence were true, then all sentences after S_i would have to be untrue, for instance S_{i+1} would have to be untrue, but that would only be the case if there were some sentence after S_{i+1} that was true, but that would contradict the assumption that all sentences after S_i are untrue. If S_i is untrue however, some sentence after S_i would have to be true, and the same problem would arise for that sentence as the one that arose if S_i itself were true. This shows that paradox can arise without self-reference, and thus that blocking self-reference is not sufficient for forming an adequate theory of truth.

1.1.4 The Empirical Liar

Kripke (1975) showed that it cannot always be determined whether or not a sentence is paradoxical based only on its syntactic form. He illustrated this with the 'Nixon-Watergate example':

Jones: "A majority of Nixon's assertions about Watergate are false." (1.1)

Nixon: "Everything Jones says about Watergate is true" (1.2)

In most cases, this set of sentences will be completely unproblematic. If, for instance, Nixon said, apart from (1.2), only a number of false things about Watergate, and Jones only true things, then (1.1) and (1.2) are both simply true. But suppose that (1.1) is Jones's only assertion about Watergate, and suppose that Nixon's assertions other than (1.2) are equally divided between true and false. Then (1.1) now claims of (1.2) that it is false, while (1.2) claims of (1.1) that it is true, and paradox ensues.

It stands to reason that a theory of truth should allow (1.1) and (1.2) to become simply true or false in the appropriate circumstances, and only be treated as paradoxical if the empirical facts make them out to be.

1.2 Solutions to the Liar

Over the years, a large number of systems, some lucid, some obscure, have been proposed to introduce truth in a language but avoid triviality due to the Liar. The most influential of these are the works of Tarski (1935) and Kripke (1975), which also bear direct relevance to the theory of Hansen. These theories and will be discussed in the next section. The section is followed by a very brief outline of some other influential theories in the literature that have less direct relevance for Hansen’s theory.

1.2.1 Towards Hansen (2014)

The theory of truth as presented by Hansen, which will be central in this thesis, is a direct reply to Kripke’s work on truth, which in turn replies to Tarski. This section describes the core of these theories to provide a background relative to which Hansen’s work must be viewed.

Tarski

The work of Tarski forms the starting point of practically every theory of truth. Many current theories reject some of his ideas about truth and his strategy for avoiding paradox is outmoded, but his early work on truth and the conclusions he has drawn about the definability of truth form a guide for anyone who embarks on forming a theory about truth. Tarski’s work was indeed groundbreaking at the time and is still referred to by almost everyone working in the field.

Ideas about truth

Tarski (1935) described what it intuitively means for a sentence to be true by his convention T or T-schema,

$$T(\text{“}\phi\text{”}) \leftrightarrow \phi,^3$$

expressing that one can claim that a sentence is true precisely when one can claim the sentence itself. Tarski argued that any suitable theory of truth should at least obey this schema.⁴

Tarski also proved with an argument similar to that for Gödel’s incompleteness theorems, that any language that contained basic arithmetic (like classic first-order logic) can not have the unrestricted T-schema because that would

³With ϕ an arbitrary formula.

⁴Tarski proposes the schema as a test for a potential theory. This does not mean that every theory that satisfies the T-schema is therefore a good theory according to Tarski.

allow for liar-like sentences to be constructed. Therefore, Tarski concluded, no language can contain its own truth predicate.

Avoiding paradox

To speak about truth but stop the Liar from trivialising the system, Tarski suggested a hierarchy of languages. None of these languages contains its own truth predicate, but each language contains a truth predicate for the previous language. Thus, the English sentence “‘Snow is white’ is true’ is true” is actually “‘‘Snow is white’ is true₁’ is true₂”. The Liar is now ill-formed: it is an attempt to speak about truth for a language in the language itself. In order to avoid Yablo’s paradox, the hierarchy must also be well-founded.

Merits

Even though many theories of truth abandon full applicability of the T-schema, the T-schema is still often used as a tool for analysing a theory. A major flaw of Tarski’s hierarchy of languages is that it suffers from expressive weakness: sentences like empirical liar sentences cannot be formulated, even though they can be perfectly harmless under the right circumstances. Also, abandoning the idea of having one truth predicate and creating a whole family seems unnatural to most.

Kripke

Kripke, in his 1975 article ‘Outline of a Theory of Truth’, addresses the Liar Paradox and seeks to improve Tarski’s treatment of the problem. He points out several flaws in Tarski’s theory and proposes his own paracomplete theory, claiming that such theories so far “almost invariably are mere suggestions, not genuine theories”⁵.

The intuitive idea

Kripke explains the intuitive idea of his theory by imagining to explain the word ‘true’ to someone who does not understand it yet. This person is not unfamiliar with the concept of truth, he just does not know how to use the word ‘true’. Therefore, this person does know he can assert “Snow is white” (and deny “Snow is green”), but he does not know what to do with the sentence “‘Snow is white’ is true”. Now, we can tell this person that “we are entitled to assert (or deny) of any sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself”⁶. This person now enters into an iterative process where at first, he could assert “Snow is white”, then also “‘Snow is white’ is true” and at the next iteration “‘‘Snow is white’ is true’ is true”, etcetera, until he has learned to assert or deny all sentences that can correctly be asserted or denied. He will never assert or deny the liar sentence, because it is not based on any non-semantic sentence. Kripke calls

⁵Kripke (1975), p698

⁶Kripke (1975), p701

such sentences ‘ungrounded’ and classifies them as *undefined*⁷.

The formal account

In the formal account of Kripke’s theory, what corresponds to the initial stage where the person cannot use the word ‘true’ yet, is a classic first order language L that contains all sentences in its domain. Then, a partially defined truth predicate $T(x)$ is added with extension S_1 and anti-extension S_2 . The language $\mathcal{L}(S_1, S_2)$ is the language where T is interpreted according to S_1 and S_2 . The function ϕ is defined as $\phi((S_1, S_2)) = (S'_1, S'_2)$ where:

S'_1 : True sentences of $\mathcal{L}(S_1, S_2)$

S'_2 : False sentences of $\mathcal{L}(S_1, S_2)$

and elements of the domain that are not sentences of $\mathcal{L}(S_1, S_2)$.

For complex sentences consisting of more than one atomic formula, Kripke employs Strong Kleene valuations. Strong Kleene evaluations basically amount to complex sentences getting the same truth values as they would in classic logic if sufficient subformula have proper truth values, and undefined otherwise. For instance, $\phi \vee \psi$ is true if ϕ is true and ψ undefined, but undefined if both ϕ and ψ are undefined.

A total order \leq on interpretations can be defined as:

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

Since ϕ is monotone with respect to \leq , ϕ has a minimal fixed point where $T(\ulcorner s \urcorner)$ is in the extension (anti-extension) of T iff s is in the extension (anti-extension) of T . Because the Liar depends only on itself, and since it will never be among the true or false sentences of any language, it will also never become a member of the extension or anti-extension of T and it becomes undefined in the minimal fixed point.

Merits

Kripke’s theory is generally accepted as a major improvement on Tarski’s theory. It allows sentences to receive a level dynamically, and does not divide the truth predicate up into many subscripted separate predicates. It allows empirical liar sentences to be formed and to receive a proper truth value if the circumstances make them unproblematic, and to become undefined if they turn out paradoxical. The theory has also some more and less pressing shortcomings, which will be discussed in section 2.1.

1.2.2 Other approaches

This section aims to provide some further context for Hansen’s theory. It briefly describes some other theories that have been proposed to define truth without paradox and gives a taste of the alternatives to the choices made by Hansen.

⁷Ungrounded sentences are undefined in the minimal fixed point. Kripke proposes a family of different truth predicates corresponding to different fixed points. Some ungrounded sentences may have truth values in other fixed points.

Paraconsistent theories

Paraconsistent theories form the flip side of paracomplete theories, a category the theories of Kripke and Hansen fall into. Where paracomplete theories posit that some sentences lack a truth value, paraconsistent theories claim that some sentences are both true and false. The paraconsistent view is most strongly advocated by Priest (1984, 2006), who created a logic known as Logic of Paradox. Unlike paracomplete theories, paraconsistent theories retain Law of Excluded Middle but reject Ex Falso Sequitur Quod Libet. The interpretation of the truth predicate can be constructed in a fashion similar to that proposed by Kripke for the paracomplete theories, only the extension and anti-extension of the truth predicate are not required to be disjoint, but instead are required to jointly exhaust the domain.

Contextualism

The contextualist approach to truth is based on the Tarskian hierarchy of languages. Advocates of contextualism are Parsons (1974) and Glanzberg (2001). Truth claims are seen as context dependent, and the truth predicate carries a contextual parameter. The contexts roughly correspond to Tarski's levels, and will also be referred to as levels. Truth claims then have a different semantic status in different contexts. The contextualist reasoning for the Liar is as follows: the Liar is in some sense ill-formed and fails to express a proposition. But then, given that fact and the meaning of the Liar, it is also actually true. Let L stand for the Strengthened Liar and let L_i be the interpretation of L at some level i . L_i is equal to $\neg T_i(L_i)$, which is neither true nor false. But at later stages one can claim $\neg T_k(L_i)$, $k > i$, which is not itself the Strengthened Liar but rather says *of* the Strengthened Liar (L_i) that it is not true, and that claim is true.

Hansen's theory is not a contextualist theory, but he does borrow from contextualism. Hansen, like Kripke, Tarski, and the contextualist approach, works with some form of levels and as for contextualism, the semantic value of a sentence in a level depends on the semantic value of other sentences at that same level. Hansen also creates a way to express the fact that the Strengthened Liar is not true, which is evaluated at a later stage than the Strengthened Liar itself.

Revision Theory

The Revision Theory of Truth was independently conceived of by Herzberger (1982) and Gupta (1982), and is described in work from Herzberger and Gupta and Belnap. Unlike most other theories of truth, Revision Theory does not block the Liar but models its semantic behaviour. And unlike most other theories of truth, Revision Theory uses a classical setting. The central idea of Revision Theory is to assign all sentences some truth value as initial hypothesis. Then, sentences are evaluated relative to that hypothesis and using the Tarski bi-conditional in a slightly adapted form to get the next hypothesis.

These sequences of subsequent hypotheses form revision sequences, and different initial hypothesis lead to different revision sequences. Sentences can then be categorised based on the behaviour of their truth value in and across revision sequences.

1.3 Research question

In this thesis, Hansen’s approach to truth and the Liar will be discussed. The properties of Hansen’s system will be studied and evaluated. The theory will be compared to Kripke’s theory on which Hansen seeks to improve. It is clear from Hansen (2014) that the system is a major achievement in defining truth without triviality and with great expressive strength while retaining most of the advantages of Kripke’s approach⁸. However, this achievement also comes at a cost and this thesis is an attempt to give an overview of the costs and benefits of Hansen’s approach. This aim can be formulated in a simple question: How well does Hansen’s theory model truth and avoid paradox?

1.4 Relevance for Artificial Intelligence

Investigating the Liar has strong ties to AI and this section aims to give an idea of how the Liar is linked to other concepts in the field of Artificial Intelligence. It will look at other paradoxes occurring in the field, the gap between humans and machines and the concept of truth. It will turn out that theories of truth could impact Artificial Intelligence on a number of different levels.

1.4.1 Paradoxes

Paradoxes occur in many fields like semantics, epistemology and set-theory. Paradoxes similar in structure to the Liar include the Russell Paradox, Cantor’s paradox, the Hypergame Paradox, the paradox of the Knower, Grelling’s paradox, Berry’s paradox, Richard’s paradox, Quine’s quotation paradox and the Brandenburg-Keisler paradox.⁹

The Knower Paradox is concerned with knowledge and of particular interest to Artificial Intelligence due to its relation to agents who may wish to reason about their own or other agents’ knowledge. The Knower Paradox can be constructed with the sentence “This sentence is not known by anyone”. If the sentence is not true, then it has to be known by someone, but if it is known by someone, it must be true. Any agent can apply this reasoning and arrive at this conclusion, thus presumably knowing “This sentence is not known by anyone”, but that would imply that the sentence is false.

⁸In particular, the Semantic Liar becomes problematic only when the empirical facts make it so.

⁹More information on these paradoxes can be found in Bolander (2014).

The circularity of the Liar paradox is also found in the definition of common knowledge, although this definition is not paradoxical. A proposition p is common knowledge between a group of agents if each agent knows that p and that p is common knowledge.

Argumentation-theory encounters the problem of self-defeating arguments, like arguments claiming of themselves that they are not correct (the Liar), or the more complex problem of a witness claiming of himself that he is unreliable. The role of self-defeat in argumentation theory is explained in more detail in Prakken (2014). In argumentation-theory, self-defeat does not lead to trivialisation of the system, but it does have some odd consequences. Argumentation-theory is of interest to Artificial Intelligence because it can be used to model the process of arriving at a particular conclusion when the evidence available is incomplete or even contradictory. Agents in real-life environments should be able to make decisions based on such evidence.

Due to the similarities between the paradoxes, a solution for one of them can guide solutions for the other paradoxes. Since the Liar Paradox is a relatively elementary version of the paradox, attempting to solve the Liar may be a good place to start. In particular, since the Knower Paradox relies on both knowledge and truth, the Liar Paradox is arguably more basic than the Knower.

1.4.2 Humans

It is a fact that humans employ the notion of truth. In everyday conversation claims to the truth or falsity of a particular claim are easily asserted. If an artificial agent is to engage in natural conversation with humans, being able to process such claims and make them when appropriate are part of an essential skill-set. This ability would require a proper theory of truth in order to avoid agents starting to exhaust smoke through their ears and shutting down whenever confronted with the Liar Paradox or retreating in infinite consideration when they happen to conceive of the sentence themselves at some unguarded moment.

Being able to deal with truth correctly would contribute to minimising the gap between humans and computers, and work towards enriching the latter with another property that is often considered typically human: the ability to reflect on their own internal state. The Stanford Encyclopedia entry on consciousness suggests as one way to define conscious creatures as creatures who are “not only aware but also aware that they are aware”. This form of ‘meta-awareness’ and self-reference has everything to do with the paradox of the Liar.

1.4.3 Truth

Apart from applications of ideas arising from theories of truth, truth is already of interest to Artificial Intelligence in its own right. The paradoxality of the Liar shows us that there is something wrong with our intuitions concerning truth and logic, and theories of truth try to unearth that mistake and replace our ideas with more consistent ones. The study of Artificial Intelligence, in the broadest

sense, can be seen as understanding the world and finding the principles that govern it. Theories of truth attempt to do just that with the concept of truth.

Although theories of truth focus mostly on languages containing their own truth predicate, the ideas about truth that originate from it do not need to be limited to such languages. For instance, Hansen does not only define the behaviour of the semantic predicates in the object-language, but also explicitly defines the truth value of sentences in the meta-language in terms of his system for truth in the object-language. Three-valued logics like the one Hansen proposes do not only provide suggestions for universal languages: they suggest alternatives for the way we conceive of truth and may even inspire researchers to question the standard of binary computers.

1.5 Structure

In this chapter the problem at hand, defining truth without creating triviality, has been introduced together with a number of different approaches that have been proposed to solve this problem. The following chapters will focus on the recent publication by Hansen, Hansen (2014). Chapter 2 gives an introduction to Hansen's theory and the process by which a number of traditionally problematic sentences receive a truth value. Also, a critical note will be placed by the definition of the system, showing that one of the principles used is redundant because it follows from another principle. The first half of the chapter is based directly on Hansen's article. The evaluations of problematic sentences come from Hansen in some cases and from the author in others. The following chapters contain new insights in Hansen's theory based on thorough analysis and are, unless otherwise indicated, the work of the author. Since Hansen attempts to improve on Kripke, chapter 3 compares Hansen's theory with Kripke's, aiming to determine to which extent the main accomplishments and defects of Kripke's theory carry over to Hansen's. Chapter 4 looks more specifically at the weaker and stronger points of Hansen's theory, although Kripke is often used for reference. The reader who is well familiar with Hansen's and Kripke's theories and wishes to get a quick idea of how to evaluate the former may wish to focus his attention on this chapter and the last, concluding chapter. The conclusion summarises the results from chapters 2-4 but does not purport to be useful without further knowledge of the previous chapters. It will also contain a verdict on the value of Hansen's system, but only in very general terms.

Chapter 2

Hansen's theory

“This is my story: There are words and there is the world, and when the former correspond to the latter, there is truth. Yet, words are in the world.” - Hansen (2014)

Kripke may be major player in the field of theories of truth and his ideas may be highly influential, his theory is not perfect. This chapter describes Hansen's system and how Hansen seeks to improve on Kripke. It will turn out that Hansen's system has a major advantage over Kripke's: it can express the undefinedness of the Liar. This chapter gives a brief version of Hansen's theory and some examples of interesting sentences and their evaluations. Also, a redundancy in the definition of the system will be proven and placed in context.

2.1 Flaws in Kripke's theory

In the literature, several points of critique on Kripke's theory have been proposed. The most pressing problem is that of expressive weakness: some sentences simply cannot be formed. A discussion of this problem is given in section 2.1.1. Also, the behaviour of compound sentences, in particular those containing implication, has been criticised. This criticism is discussed in section 2.1.2. Hansen additionally criticises Kripke's use of names and argues that another form of naming, naming using quotation marks, should be introduced. This is content of section 2.1.3.

2.1.1 Undefinedness

In Kripke's theory, the Liar is undefined, but this cannot be expressed in the theory, that is:

$$\text{The Liar is undefined} \tag{2.1}$$

cannot be formulated in Kripke's system, not only because there is no undefinedness predicate, but also because the undefinedness of the Liar is a 'meta-insight' in the theory. At each level, more sentences become true or false (and cease to be undefined), and the sentences that are undefined in the minimal fixed point are to be thought of as 'the' undefined sentences. But (2.1) cannot exist in the fixed point, because it would require the Liar to be declared undefined at a previous level, but this only happens in the fixed point itself, so if (2.1) were a part of the language, it would become true at the level after the level of the minimal fixed point, which would imply the supposed fixed point was never a fixed point to start with. This protects Kripke from the Revenge Liar, because

$$\text{The Revenge Liar: The Revenge Liar is false or undefined} \quad (2.2)$$

cannot be formulated. However, it also blocks the formulation and evaluation of intuitively true sentences like (2.1), which can be considered a flaw. Hansen adds expressive strength to the system by introducing an undefinedness predicate and formulating rules that allow sentences to become *undefined* earlier in the process.

2.1.2 Connectives

Even though Hansen himself only explicitly criticises Kripke's theory for the inexpressibility of the Liar, he also deviates from Kripke in his valuations of complex formulas. In particular, Kripke uses Strong Kleene valuations where the negation of an undefined formula is itself undefined. Hansen argues that this is not intuitive, using the sentence

$$\text{It is not the case that the Liar is false} \quad (2.3)$$

Since the Liar is undefined, it intuitively is not false indeed, so (2.3) is intuitively true. This is not consistent with choice negation used by Kripke. Hansen uses exclusion negation instead, making the negation of an undefined formula true, which makes (2.3) also true. Section 4.3 of this thesis will show that this example is not in fact an argument for the use of exclusion negation.

Hansen also uses a different truth table for disjunction, treating undefined as false. His different treatment of the truth value 'undefined' can be explained by considering the fact that Hansen uses 'undefined' as "is never going to be true or false" whereas Kripke uses it as "is not yet true or false" during the construction of the truth predicate and only as "is never going to be true or false" once that construction is finished. This difference will become more clear in the next section. Whether or not it is an improvement of Hansen's system to abandon Strong Kleene valuations and use another scheme will be discussed in Section 4.4.

Kripke's system is criticised for its implication¹, which does not satisfy $\models \phi \rightarrow \phi$. As will be discussed later in section 3.3.3, this does not hold in Hansen's system either.

¹Visser(1989), Beall & Glanzberg (2014)

2.1.3 Names

Kripke uses constants to refer directly to sentences², but Hansen argues that this does not account for the intuitive difference in natural language between

(2.5) is false (2.4)

(2.4) is true (2.5)

and

(2.7) is false (2.6)

“(2.7) is false” is true. (2.7)

The difference, Hansen argues, is that for sentences (2.4) and (2.5), there is no intuitive order in which they should be evaluated, but for (2.6) and (2.7) it is intuitive to evaluate (2.6) before (2.7), since (2.7) quotes (2.6). Hansen further argues that quotation-names always exist in natural language, and self-quotation is impossible. Therefore, quotation names have different properties from constant names, and should be treated differently by a formal theory. Kripke's language does not contain a quotation device, a gap Hansen fills by adding quotation and describing its behaviour.

2.2 Hansen's theory

This section introduces Hansen's theory, first giving the intuitive account to provide some insight into the idea of the theory and the reasons behind certain choices and then giving a more formal account of how the system is defined. This section aims to give the reader a basic understanding of the theory discussed in the rest of this thesis and explain concepts and notation used elsewhere. For a more complete account, the reader is referred to Hansen's original description in Hansen (2014).

2.2.1 Intuitive account

Hansen uses a metaphor to introduce his own theory that is similar to the one Beall³ uses to describe Kripke's theory. Kripke uses the metaphor of explaining the word 'true' to someone who does not know it yet and this person's increasing understanding, whereas Beall and Hansen use the metaphor of a writer who writes increasingly many sentences in books. The writer, in Hansen's story, writes three books: *The True*, *The False* and *The Undefined*. Initially, he just writes all non-semantic facts in *The True* and *The False*. As a result of this, more sentences become true and false, if one thinks of predicating a semantic

²Alternatively, one could define a predicate that is true of a unique sentence

³Beall (2007)

value to a sentence as claiming it is written in the corresponding book, and the writer can continue writing. This is not all that happens, however, because the writer also has *The Undefined* to write. That is, sentences are not automatically undefined as long as they do not appear in *The True* or *The False*, but they are explicitly declared *undefined* when it becomes clear it will never appear in *The True* or *The False*. How does this become clear? Hansen adopts 'three principles of hope': if none of those principles is satisfied for a sentence, there is no longer any hope of the sentence ever becoming true or false and it is written in *The Undefined*.

1. As long as more and more of the sentences on which s depends are getting truth values, there is still hope for s .
2. If there is a chance that the truth value of s can be determined without getting into circularities and infinite sequences, there is still hope for s .
3. If s quotes a sentence and the truth value of s depends on that sentence, there is still hope for s .

The list above is intended only to give an intuitive idea of what 'hope' consists in in this context, a more formal description will be provided in the next section.

2.2.2 Formal theory

Hansen distinguishes between four semantic values: *undetermined* ($()$), *undefined* ($+$), *true* (\top) and *false* (\perp). Of those, he only refers to undefined, true and false as 'truth values' and only to true and false as 'proper truth values'.

The syntax of the object language used in Hansen comes with few surprises given the considerations mentioned earlier. It mostly corresponds to classic first order logic, with four exceptions. First of all, a quoted wff is a term. Second, third and fourth he introduces truth, falsity and undefinedness predicates.

Hansen builds up his semantics through levels in the form of evaluations $\mathcal{E} = (\mathcal{T}, \mathcal{F}, \mathcal{U})$, where sentences in \mathcal{T} are true, those in \mathcal{F} false and those in \mathcal{U} undefined. Initially, all sentences are undetermined, but, Hansen proves, the iterative process of giving sentences truth values leads to a unique total and consistent evaluation where all sentences are a member of exactly one of the three sets. The levels are constructed in two stages, first of all there is the tentative evaluation with respect to the previous level ($\text{TE}_{\mathcal{E}}$) where the sets \mathcal{T} and \mathcal{F} are expanded to $\mathcal{T}_{\mathcal{E}}$ and $\mathcal{F}_{\mathcal{E}}$, and next comes the evaluation with respect to the previous level ($\text{E}_{\mathcal{E}}$), where \mathcal{U} is expanded also.

To construct the tentative evaluation, Hansen gives eleven rules, E1-E11 that reflect the truth tables in figure 2.1. Atomic sentences not containing semantic predicates are evaluated in the usual, classical way. For sentences s predicating truth, falsehood or undefinedness to a sentence s' that has already received a truth value at an earlier level, the sentence s is *true* if it has not yet

$\neg\phi$	\perp	$\phi \vee \psi$	\top	\perp	\top	\perp	\top
\top	\perp	\top	\top	\top	\top	\top	\top
ϕ	\perp	ϕ	\perp	\top	\perp	\perp	\perp
\perp	\top	\perp	\top	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp

$\phi \wedge \psi$	\top	\perp	\perp	\perp	\perp	\perp	\perp
\top	\top	\perp	\perp	\perp	\perp	\perp	\perp
ϕ	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp

$\phi \rightarrow \psi$	\top	\perp	\perp	\perp	\perp	\perp	\perp
\top	\top	\perp	\perp	\perp	\perp	\perp	\perp
ϕ	\perp	\top	\top	\top	\top	\top	\top
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp

$\phi \leftrightarrow \psi$	\top	\perp	\perp	\perp	\perp	\perp	\perp
\top	\top	\perp	\perp	\perp	\perp	\perp	\perp
ϕ	\perp	\perp	\top	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp
\perp	\perp	\perp	\perp	\perp	\perp	\perp	\perp

Figure 2.1: Truth tables

been declared undefined and s' is true, false or undefined respectively. If s' has another truth value, s is false.

To construct the evaluation, two new concepts need to be introduced. First and foremost, the notion of *dependency*. Hansen defines:

The binary relation $R_{\mathcal{E}}$ on \mathcal{S}^4 , called the *direct dependency relation with respect to the evaluation \mathcal{E}* , is defined as follows: $sR_{\mathcal{E}}s'$ if both s and s' are undetermined according to \mathcal{E} , and s is $\neg\phi$ and s' is ϕ , or s is $\phi \vee \psi$, $\phi \wedge \psi$, $\phi \rightarrow \psi$ or $\phi \leftrightarrow \psi$ and s' is ϕ or ψ or s is $\exists v\phi$ or $\forall v\phi$ and s' is $\phi(v/c)$ for some constant c , or s is $T(t)$, $F(t)$ or $U(t)$ and s' is $I(t)$.⁵
Hansen (2014), p227

Dependency ($\overline{R}_{\mathcal{E}}$), is the transitive closure of $R_{\mathcal{E}}$. Also, $\overline{R}_{\mathcal{E}}(s) := \{s' | s\overline{R}_{\mathcal{E}}s'\}$. Secondly, Hansen defines $\mathcal{E}|\mathcal{S} := (\mathcal{T} \cap \mathcal{S}, \mathcal{F} \cap \mathcal{S}, \mathcal{U} \cap \mathcal{S})$.

Now, the sentences that become a member of $\mathcal{U}_{\mathcal{E}}$ are the sentences s such that each of the following holds:

- E1) $s \notin \mathcal{T}_{\mathcal{E}} \cup \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}$
- E2) $\text{TE}_{\mathcal{E}}|\overline{R}_{\mathcal{E}}(s) = \mathcal{E}|\overline{R}_{\mathcal{E}}(s)$,
- E3) for every sentence s_0 , if $s\overline{R}_{\mathcal{E}}s_0$ then ($s_0\overline{R}_{\mathcal{E}}s$ or there is an infinite $\overline{R}_{\mathcal{E}}$ -sequence $s_0\overline{R}_{\mathcal{E}}s_1\overline{R}_{\mathcal{E}}s_2\overline{R}_{\mathcal{E}}\dots$ consisting of distinct elements), and
- E4) there is no sentence s' which is quoted in s such that $s\overline{R}_{\mathcal{E}}s'$.

E1 guarantees that if a sentence has received a truth value at a particular level, it does not get re-evaluated at a later level. That this is relevant will be discussed later in section 4.5. E2-E4 correspond directly to the three principles

of hope. E2 states that if some sentence s depends on has received a proper truth value in the last tentative evaluation, it will not become undefined. E3 states that a sentence only becomes undefined if all sentences it depends on either depend on s (circularity) or are the beginning of an infinite dependency chain (as is the case for the sentences in Yablo's paradox). In section 2.5 it will be shown that E2 follows from E3. E4 is related to the supposed intuition that a sentence is always only evaluated after whatever sentences it quotes is evaluated. How this works out is shown with some examples in section 2.4.3. Criticism on E4 and an alternative criterion are given in section 4.9.

This concludes the presentation of Hansen's formal theory for this thesis. The essential concepts have been introduced and this paragraph should give a firm background for understanding discussions of specific aspects of Hansen's theory. Also, the words 'semantic value', 'truth value', 'proper truth value' and 'dependency' when used in this thesis will refer to the concepts as defined by Hansen.

2.3 Clarifications

When Hansen describes his system, he leaves some things unsaid. The first question raised is how exactly being *quoted in* is defined. Hansen is not very explicit about this, but there is a fairly obvious answer to this question. The second, more opaque issue is the issue of α -equivalence. It is not clear whether or not Hansen intended α -equivalence to hold in his system.

2.3.1 Quoted in

Hansen defines what it means for a sentence ϕ' to be *quoted in* ϕ that $\ulcorner \phi' \urcorner$ appears in ϕ , and ϕ' is equal to ϕ'' except that all present free variable occurrences in ϕ'' are replaced with constants.

This begs the question what 'appears in' signifies in the context. It could mean that the symbol string $\ulcorner \phi' \urcorner$ (with ϕ'' replaced with a specific formula) is a part of the string for ϕ , or that ϕ alternatively could contain the name of some formula that quoted ϕ' .

The way Hansen treats the sentences (2.6) and (2.7) suggests that the intended interpretation is the former, not the latter. Hansen claims that (2.6) becomes undefined, but if (2.6) quoted any sentence that was quoted in a sentence (2.6) contained the name of, it would quote the sentence "(2.7) is false", that is: it would quote itself, and it would not become undefined. What is more, because it would forever depend on itself and forever quote itself, it would never receive a truth value at all, which by no means can be what Hansen intended.

If one instead assumes that ϕ quotes ϕ' if and only if the symbol string $\ulcorner \phi' \urcorner$ (with ϕ'' replaced with a specific formula) is a part of the string for ϕ and ϕ' is equal to ϕ'' except that all present free variable occurrences in ϕ'' are replaced with constants, then only (2.7) would still quote another sentence, and (2.6) and (2.7) would become undefined and false, as Hansen intended.

2.3.2 α -equivalence

When are two sentences the same? If, for instance, a constant c_s is set to refer to the sentence $\forall x(P(x) \rightarrow \neg T(x))$, does it also refer to $\forall y(P(y) \rightarrow \neg T(y))$? Hansen does not mention α -equivalence and it is not clear what his intentions with regard to this topic are. In the following discussion, s and s' will be used to refer to the two sentences in this paragraph.

With constants, two sentences that are the same except for co-referring constants, for instance $\neg T(c_{sl})$ and $\neg T(c_{nsl})$, with $I(c_{sl}) = I(c_{nsl}) = \neg T(c_{sl})$, can have different truth values because one ($\neg T(c_{sl})$) is self-referential and the other ($\neg T(c_{nsl})$) is not. For these sentences, the first is undefined and the second is true.

This effect can be replicated with quantified formulas s and s' . Let $I(P) = s$, and let α -equivalence not hold for identity. Then, s is the sentence that says of itself that it is not true, and s' is the sentence that says of s that it is not true. If α -equivalence were to hold, both sentences would say of themselves that they are not true. This issue is addressed again in section 3.3.

Note that absence of α -equivalence is not necessary to express the fact that s is not true. This could simply be expressed with the sentence $\neg T(c_s)$, which would evaluate to true. Thus, banning α -equivalence from identity is most similar to the treatment of constants, but it is not necessary to express semantic facts.

2.4 Undefined sentences

This section gives examples of sentences that become undefined and the process of their evaluation. It is intended to give some more insight into the workings of the system as defined by Hansen, as well as showing the results it gives for some tricky sentences.

2.4.1 Undefined at higher levels

The example of the Liar may make the impression that sentences only become undefined at the first level. After all, if a sentence is circular, it is circular from the start. In a sense, that is true, but consider the sentence

$$F(c_{2.8}) \wedge T(c_b) \tag{2.8}$$

where $I(c_{2.8}) = F(c_{2.8}) \wedge T(c_b)$.

Now, if c_b refers to some true formula like $0 = 0$, then $T(c_b)$ will also become true and the truth value of the sentence will come to depend only on itself and it will become undefined. If c_b refers to some false or undefined formula like $0 = 1$, $T(c_b)$ will become false and therefore the whole formula becomes false: the truth value of $F(c_{2.8})$ does not matter anymore. But this information, whether $T(c_b)$ is true, false or undefined, is not yet available at E_{E^0} . If $I(c_b) = (0 = 0)$,

then $0 = 0$ becomes true at TE_{E^0} , and thus E2 fails for $T(c_b)$ and $F(c_{2.8})$ at E_{E^0} . This process repeats itself at E^2 where $T(c_b)$ becomes true and E2 fails for $F(c_{2.8})$. Only at E_{E^2} ($= E^3$) does $F(c_{2.8}) \wedge T(c_b)$ become undefined. This event can be delayed as long as one likes by adding extra truth predicates to the second conjunct.

2.4.2 The Disjunctive Liar

The Disjunctive Liar is a version of the Liar where the claim that this sentence is false is disjuncted with some falsehood, for instance

$$F(c_d) \vee 0 = 1 \quad (2.9)$$

where $I(c_d) = F(c_d) \vee 0 = 1$. At TE_{E^0} , $0 = 1$ becomes false, so at E_{E^0} E2 fails for $F(c_d)$. At the next level however, $F(c_d) \vee 0 = 1$ only still depends on itself and at E_{E^1} , it becomes undefined. Note that at that point, $F(c_d)$ depends only on $F(c_d) \vee 0 = 1$ and $F(c_d) \vee 0 = 1$ depends only on $F(c_d)$, so E3 is satisfied for $F(c_d)$ and since E1, E2 and E4 are too, $F(c_d)$ becomes undefined at the same level.

2.4.3 Sentences containing quotes

There are two types of sentences that contain quotes and satisfy E3: sentences that satisfy E3 because at some point they only depend on the quoted sentence(s) (and all sentences that sentence depends on), and sentences that satisfy E3 because at some point they only depend on sentences that are not quoted (and all sentences those sentences depend on).

An example of the first category is the sentence $0 = 0 \wedge T(T(c))$, where c refers to the sentence itself. At the first level, $0 = 0$ becomes true and the sentence now only depends on itself and sentences that depend on the sentence itself, but because E4 fails, it does not become undefined. Instead, $T(c)$ becomes undefined (at level 2) and $T(T(c))$ becomes false as a result, which causes the whole formula to be false. Sentences from this category never become undefined: the only source of self-reference or infinite dependency is inside the quoted sentence, but since by E4 the sentence always waits until the quoted sentence has received a truth value, the sentence always receives a proper truth value.

An example of the second category is the sentence $T(0 = 0) \wedge T(c)$, where c refers to the sentence itself. At TE_{E^0} , $0 = 0$ becomes true. At the next level, because $0 = 0$ has a truth value, $T(0 = 0) \wedge T(c)$ does not depend on it anymore, so E4 is satisfied for $T(0 = 0) \wedge T(c)$ at E_{E^1} . However, because $T(0 = 0)$ receives a proper truth value at TE_{E^1} , E2 is not satisfied. At level 3, no formulas receive a proper truth value, but E4 is still satisfied for $T(0 = 0) \wedge T(c)$ and because the only conjunct it now depends on is $T(c)$, and $T(c)$ depends on $T(0 = 0) \wedge T(c)$, E3 is now satisfied for $T(0 = 0) \wedge T(c)$ and it becomes undefined.

2.4.4 Undefined subformulas

Formulas with undefined subformulas can themselves have proper truth values. For instance, the disjunction of the Liar and the Truth Teller,

$$F(c_l) \vee T(c_t) \tag{2.10}$$

where $I(c_l) = F(c_l)$ and $I(c_t) = T(c_t)$, is false. Both the Liar and the Truth Teller become undefined at the first level. The disjunction depends on both the Liar and the Truth Teller, but neither formula depends on the disjunction. Therefore, E3 is not satisfied for $F(c_l) \vee T(c_t)$ at the first level. At TE_{E^1} , both disjuncts have a truth value and by TE3 $F(c_l) \vee T(c_t)$ can receive a proper truth value: false.

2.4.5 Satisfying E3

If a sentence depends on another sentence that does not satisfy E3 and therefore does not become undefined, the sentence itself does not satisfy E3 either and also does not become undefined. This property is stated in the following theorem.

Theorem 1. *For sentences s and s' , if $s \overline{R}_{E^\alpha} s'$ E3 is not satisfied for s' at E_{E^α} , then it is not for s either.*

Proof. If E3 failed for for s' , then there must have been a sentence s_0 with $s' \overline{R}_{E^\alpha} s_0$, such that neither $s_0 \overline{R}_{E^\alpha} s'$, nor was there an infinite sequence $s_0 \overline{R}_{E^\alpha} s_1 \overline{R}_{E^\alpha} s_2 \overline{R}_{E^\alpha} \dots$ consisting of distinct elements. s_0 also causes E3 to fail for s . Suppose for the sake of contradiction that $s_0 \overline{R}_{E^\alpha} s$ to make E3 hold for s . But because $s \overline{R}_{E^\alpha} s'$, by transitivity, $s_0 \overline{R}_{E^\alpha} s'$, which contradicts our assumptions. Therefore, s_0 does not depend on s . s_0 is also not the beginning of an infinite dependency chain. Because $s \overline{R}_{E^\alpha} s_0$ (transitivity), but neither $s_0 \overline{R}_{E^\alpha} s$ nor is s_0 the beginning of an infinite dependency chain consisting of distinct elements, E3 fails for s if E3 fails for s' . \square

2.4.6 Undefined formulas and subformulas

Sentences that become undefined become undefined together with all its undetermined subformulas.

Theorem 2. *If a sentence s has subformulas s_1, s_2, \dots, s_n of which formulas s_1, s_2, \dots, s_k ($k \leq n$) are undetermined at E_{E^α} , and s becomes undefined at E_{E^α} , then s_1, s_2, \dots, s_k also become undefined at E_{E^α} .*

Proof. This theorem amounts to claiming that if E1-E4 are satisfied for s , then they are satisfied for all undetermined subformula of s , s_1, s_2, \dots, s_k .

- s_1, s_2, \dots, s_k are assumed to be undetermined at the beginning of E_{E^α} , so E1 must have been satisfied for s_1, s_2, \dots, s_k .

- Theorem 3 shows that the case for E2 follows from the case for E3.
- Theorem 1 shows that if there were an element of s_1, s_2, \dots, s_k for which E3 was not satisfied, it would not have been for s either. But E3 is assumed to be satisfied for s , so such an element can not exist.
- If s does not depend on any sentence it quotes then all sentences that are quoted in s have received a truth value. Because s_1, s_2, \dots, s_k are subformula of s , the same must hold for s_1, s_2, \dots, s_k .

□

2.4.7 Revenge

Consider the formalisation of the Strengthened Liar:

$$\neg T(c_{sl})$$

where $I(c_{sl}) = \neg T(c_{sl})$. The Strengthened Liar is clearly self-referential. At E_{E^0} , it satisfies all of E1-E4 and becomes undefined. But being undefined means not being true, which usually reinstates paradox in paracomplete logics. For Hansen, it does not. After the Strengthened Liar is made undefined, it is not re-evaluated (which would make it true). Hansen justifies this by arguing that sentences are undefined when the sentence that would decide its truth value cannot receive a truth value before the sentence itself does. This is the case for the Strengthened Liar, and the result is that it is undefined, and only undefined: it is not also true.

2.4.8 Self-reference

In the last section it was explained why the Strengthened Liar is undefined and only undefined in Hansen's theory. However, there is a sense in which the Strengthened Liar is indeed not true. This can be expressed using the sentence:

$$\neg T(c_{nsl}) \tag{2.11}$$

where $I(c_{nsl}) = \neg T(c_{sl})$. Even though the Strengthened Liar and (2.11) are of the same form and their constants refer to the same formula, they are not the same. The Strengthened Liar refers to itself, but (2.11) refers to another sentence, the Strengthened Liar, and claims of that sentence that it is not true. Since the Strengthened Liar is undefined, it is indeed not true and (2.11) becomes true at TE_{E^1} .

2.4.9 Yablo's paradox

Hansen formalises the Yablo sequence as: for each $n \in \mathbb{N}$ the formalisation of Y_n is

$$\forall x(P(\bar{n}, x) \rightarrow \neg T(x))$$

with $I(\bar{n}) = n$ and $I(P) = \{(n, Ym) \mid n, m \in \mathbb{N} \text{ and } m > n\}$.

It is clear that none of these sentences becomes true or false at TE_{E^0} , because they all still depend on all the following sentences to receive a truth value, and no sentence is the last one. It is interesting to see what sentences the Yablo sentences exactly depend on. Take for instance Y3. Y3 is defined as $\forall x(P(\bar{3}, x) \rightarrow \neg T(x))$. Due to the universal quantification, Y3 depends on all instances of $P(\bar{3}, x) \rightarrow \neg T(x)$, for instance the instantiation with (x/c_2) , where $I(c_2) = Y2$. Since at E^0 , everything is undetermined and since the dependency relation is defined relative to the previous level, throughout E^1 , the mentioned instantiation of Y3 depends on $\neg T(c_2)$, and therefore also on c_2 and everything c_2 depends on. This reasoning shows that throughout E^1 , all Yablo sentences depend on all sentences of the form $P(\bar{n}, x) \rightarrow \neg T(x)$, including those that are instances of Yablo sentences earlier in the sequences, and including those that are not instantiations of Yablo sentences in the first place, like $P(\bar{5}, c_2) \rightarrow \neg T(c_2)$. For many of those sentences, $P(x, y)$ is false and the whole sentence is true. The effect of this is that at E_{E^0} ($= E^1$), E2 is not satisfied for any of the Yablo sentences and none of them becomes undefined. This brings the process to TE_{E^1} , where all Yablo sentences only still depend on the Yablo sentences further down in the sequence, but due to the infinity of the sequence, they can still not receive a proper truth value. At E_{E^1} ($= E^2$) however, E1-E4 are satisfied for all Yablo sentences and they all become undefined.

2.5 E2

This section concerns the definition Hansen gives for his system, in particular rule E2. It turns out that E2 is not necessary to obtain the behaviour specified by the system, because E2 follows from E3. This insight will simplify the definition of the system and make it easier to prove certain properties. However, E2 still has its use: it may be true that if E2 fails, E3 necessarily fails also, but if both fail it is often easier to see that E2 fails. E2 is a very straightforward condition, that is, especially when doing manual evaluations of sentences, it is easy to check whether or not it is satisfied. E2 will therefore continue to be referenced in the remainder of this thesis, but will not be dealt with separately in proofs.

Note that by definition of TE1-TE11, a sentence s receives a proper truth value at some level α only if $\bar{R}_{E^{\alpha-1}} = \emptyset$, or if some other sentence it depends on receives a proper truth value at that level and this is enough to determine a proper truth value for s . In the first case, the sentence will be said to receive a truth value by itself at level α .

Theorem 3. *For all sentences s , if s satisfies E3 at some level $\alpha + 1$, it also satisfies E2.*

Proof. Assume that s satisfies E3 at E_{E^α} . The aim is to prove that

$$TE_{E^\alpha}|\overline{R}_{E^\alpha}(s) = E^\alpha|\overline{R}_{E^\alpha}(s),$$

that is:

for all $s' \in \overline{R}_{E^\alpha}(s)$, if $s' \notin E^\alpha$ then $s' \notin TE_{E^\alpha}$.

Let Σ be the set of sentences which are in $\overline{R}_{E^\alpha}(s)$ and receive⁶ a truth value at TE_{E^α} . We will prove that if Σ is non-empty, it is required that another sentence which is in $\overline{R}_{E^\alpha}(s)$ but is not in Σ also receives a truth value at TE_{E^α} , which contradicts the stipulation that Σ contains precisely those sentences s depends on which receive a truth value at TE_{E^α} . Therefore, Σ must be empty and the theorem holds.

Consider that since E3 holds for s , for all $s' \in \overline{R}_{E^\alpha}(s)$, either

1. $s' \overline{R}_{E^\alpha} s$, or
2. there is an infinite \overline{R}_{E^α} -sequence $s' \overline{R}_{E^\alpha} s_1 \overline{R}_{E^\alpha} s_2 \overline{R}_{E^\alpha} \dots$ consisting of distinct elements.

Because Σ is a subset of $\overline{R}_{E^\alpha}(s)$, the same holds for all elements of Σ .

Assume Σ to be non-empty and let σ be an arbitrary element of Σ . We will show that σ can not receive a proper truth value at TE_{E^α} unless some element outside Σ does. Either (1) or (2) must hold for σ . If (1) holds for σ , it did not receive a proper truth value by itself at TE_{E^α} since it still depended on s . Therefore, if (1) holds for σ , σ can only have received a proper truth value at TE_{E^α} if another sentence it depended on did so. If (2) holds for σ , it did not receive a proper truth value by itself because it still depended on the next element in the infinite dependency chain it was part of. Therefore, if (2) holds for σ , σ can only have received a proper truth value at TE_{E^α} if another sentence it depended on did so.

This shows that for any element of Σ to receive a truth value at TE_{E^α} , it is required that another sentence does so too. Because this is the only way any element of Σ can receive a truth value at this tentative evaluation, none of the elements of Σ will receive truth value at TE_{E^α} , unless there is at least one sentence $\overline{\sigma}$ outside of Σ that receives a truth value and causes one or more elements of Σ to receive a proper truth value. Note that because at least one element of Σ depends on $\overline{\sigma}$ and s depends on all elements of Σ , by transitivity of the dependency relation s depends on $\overline{\sigma}$ also. Since Σ was stipulated to be the set of elements which receive a proper truth value at TE_{E^α} , if Σ is not empty, such a sentence $\overline{\sigma}$ must exist to make it possible for the elements of Σ to receive a truth value. At the same time, it cannot exist because any element s depends on that receives a proper truth value at TE_{E^α} must be in Σ . This shows that the assumption that Σ is non-empty leads to a contradiction, and so Σ must be empty. \square

⁶The term 'receive' will be used here to refer solely to newly receiving, not to maintaining. If a sentence first satisfies TE1-TE11 at some level α , it will still satisfy those conditions at level $\alpha + 1$, but it will only be said to receive a truth value at level α .

Theorem 3 shows that E2 is redundant for all levels, starting from level 1. Because E1-E4 are only used to construct evaluations for succes ordinals greater than 0, this is enough to show that E2 is redundant for the system.

E2 and E3 are not equivalent. Consider the sentence $T(c_{nl})$ where $I(c_{nl}) = F(c_l)$ where $I(c_l) = F(c_l)$. $T(c_{nl})$ depends only on sentences that do not receive a truth value at TE_{E^0} , so E2 is satisfied, but E3 is not satisfied because $T(c_{nl})$ depends on $F(c_l)$, but $F(c_l)$ does not depend on $T(c_{nl})$ or a sequence of infinitely many distinct elements. This shows that E2 can be satisfied without E3 being satisfied.

2.6 Conclusion

Kripke's theory of truth suffers from a number of defects, most importantly it suffers from expressive weakness: the concept of undefinedness is not modelled in the system and cannot be modelled in the system as it is. Hansen solves this problem by devising a method to explicitly declare sentences undefined during the iterative process of evaluation using four conditions that determine if a sentence must be declared undefined. He therefore introduces, apart from a truth predicate, also a falsity and undefinedness predicate.

There is also criticism for the behaviour of the implication in Kripke's system. The behaviour of the implication in Kripke's and Hansen's systems will be discussed in section 3.3.3. Hansen also criticises Kripke's definition of negation and suggests another definition. This difference will be discussed in section 4.3.

Hansen makes an extra distinction in types of names a sentence can have. Apart from constants that refer to sentences, one can also use quotes to create a name of a sentence. This corresponds to a common sentence-naming practice in natural language and also has some special properties.

Some effects of these alterations have been discussed in section (2.4), which also further illustrates the evaluation process. It turns out that Hansen's second rule to govern undefinedness, E2, follows from E3 and is redundant. This is the content of section (2.5).

Chapter 3

Hansen and Kripke

“We make utterances which we hope will turn out to be grounded” - Kripke (1975)

Hansen intended his theory to be an improvement on Kripke’s theory, but how do the two compare? Does Hansen retain the favourable results of Kripke’s theory? The first section of this chapter shows that the sentences that receive proper truth values in Kripke’s system receive the same truth value in Hansen’s system. The second section looks at Kripke’s other fixed points and how this idea can be replicated for Hansen. Next, a series of semantic equivalences is considered and the behaviour of both systems with respect to those equivalences is analysed. The chapter has a close relationship with the next chapter, which focusses more on the value of Hansen’s system but often uses the results of Kripke’s system for comparison.

3.1 Kripke’s grounded sentences

Kripke defines ‘groundedness’ in a more narrow sense than Hansen, as Hansen allows sentences to be grounded in the ungroundedness of other sentences and Kripke does not allow for this construction. Kripke informally describes his notion of groundedness, which we will call K-groundedness, as:

In general, if a sentence such as (1)¹ asserts that (all, some, most, etc.) of the sentences of a certain class C are true, its truth value can be ascertained if the truth values of the sentences in the class C are ascertained. If some of these sentences themselves involve the notion of truth, their truth value in turn must be ascertained by looking at other sentences, and so on. If ultimately this process terminates in sentences not mentioning the concept of truth, so that the truth value of the original statement can be ascertained, we call the original sentence grounded; otherwise, ungrounded.

Kripke (1975) p693

Intuitively, K-grounded sentences are certainly true or false and no proper theory of truth should give them any other truth value. The aim is to prove that Hansens theory does in fact do this by proving that sentences that receive a proper truth value in Kripke's minimal fixed point receive the same truth value in Hansens total evaluation.

3.1.1 Technical requirements

To be able to say that the same sentence is evaluated in both systems, the symbol string that the sentence consists of needs to be a sentence in both systems. From the definitions of syntax Kripke and Hansen give, it is easily seen that all K-sentences are also H-sentences. Not all H-sentences are K-sentences, since Hansen allows for quotation and the use of falsity and undefinedness predicates. *We will call a sentence a proper K-sentence when it is itself a K-sentence and depends only on K-sentences.* What we want to prove is that all proper K-sentences that evaluate to true or false in Kripke evaluate to the same truth-value in Hansen.

Since sentences are evaluated relative to a model, it makes sense to prove the equivalence in evaluation for Kripke and Hansen relative to the same model. For this to be possible, the model has to satisfy both the conditions imposed on models by Kripke, and the conditions imposed by Hansen.

The only explicit constraint Kripke poses on the interpretation function is that it interprets all primitive (Hansen: ordinary) predicates. Hansen additionally requires that all constants have an interpretation in the domain (something Kripke may also assume), that there is a constant for each object in the domain, and stipulates that the interpretation of a quotation-name is the sentence it quotes.

For the domain, both Kripke and Hansen require that the domain at least contains all sentences in their language. As argued above, all K-sentences are also H-sentences, so any H-domain is automatically also a K-domain.

Since Hansen imposes the same constraints on both domains and interpretation functions as Kripke and then some more, it follows that every H-model is also a K-model, so we require of the model relative to which sentences are evaluated simply that it is a H-model.

3.1.2 Sentences do not become undefined too early

In order to prove the desired equivalence, a property of Hansens system will be proven first that will turn out to be helpful in proving the equivalence in evaluations.

Theorem 4. *If a proper K-sentence s depends on a possibly infinite set of sentences $s_1, s_2 \dots$ and $s_1, s_2 \dots$ receive proper truth values at level α with α a successor ordinal, then s is not undefined at level $\alpha - 1$.*

Proof. For s_k to receive a proper truth value at α , it cannot have been undefined at levels 0 up to and including $\alpha - 1$. If s_k is not undefined at level $\alpha - 1$, then s_k did not satisfy all of E1-E4 at any level β between zero and $\alpha - 1$. The following shows that if for all s_i , any of E1-E4 was not satisfied at β , then it was not for s either.

- If E1 was not satisfied for some s_k , s_k was an element of $\mathcal{T}_\beta \cup \mathcal{F}_\beta \cup \mathcal{U}$, which contradicts our assumption that all s_i receive a proper truth value at level α . Therefore, E1 must have held for all s_i .
- Theorem 3 shows that the case for E2 follows from the case for E3.
- This holds by Theorem 1 and the observation that proper K-sentences are a subset of the H-sentences.
- If E4 was not satisfied for some s_k , s_k must have quoted some sentence, which means that s_k is not a K-sentence, which means that s is not a proper K-sentence, which contradicts our assumptions. Therefore, E4 must have held for all s_i .

This shows that E1-E4 could not all have been satisfied for s at levels 0 through $\alpha - 1$, and s is not undefined at level $\alpha - 1$ \square

Note that we require *all* sentences s depends on to receive a proper truth value at level α . s may have previously depended on sentences that received a truth value in earlier levels than α , but by the definition of dependency, as soon as they received a truth value, s ceased to depend on them.

3.1.3 K-grounded sentences and proper truth values

We can now state the desired equivalence:

Theorem 5. *For every H-model \mathcal{M} and every proper K-sentence s , if s is in the extension (anti-extension) of T in the Kripke minimal fixed point, it is in the set \mathcal{T} (\mathcal{F}) of the Hansen total evaluation.*

Proof. In the following, unless otherwise indicated, ‘sentence’ is used to mean ‘proper K-sentence’.

For s to be in the extension (anti-extension) of T in the minimal fixed point, there must be some level α where s first became a member of the extension (anti-extension) of T . We will prove by induction that for all levels α , if a sentence becomes a member of the extension (anti-extension) of T , it also becomes a member of the set \mathcal{T} (\mathcal{F}) at that level. If s becomes a member of \mathcal{T} (\mathcal{F}) at some level α , by monotonicity it must also be a member of \mathcal{T} (\mathcal{F}) in the total evaluation.

The induction start is trivial as $\mathcal{L}^0 = \mathcal{L}(\emptyset, \emptyset)$ and $E^0 = (\emptyset, \emptyset, \emptyset)$.

In the induction step, first take the case of $\mathcal{L}^\alpha = \phi(S_1^{\alpha-1}, S_2^{\alpha-1})$ and $TE_{E^{\alpha-1}} = (\mathcal{T}^\alpha, \mathcal{F}^\alpha, \mathcal{U}^{\alpha-1})$, where α is a successor ordinal. ϕ is the function that takes a language to the next level as defined by Kripke.

By induction, all sentences that are in $S_1^{\alpha-1}$ are in $\mathcal{T}^{\alpha-1}$ and all sentences that are in $S_2^{\alpha-1}$ are in $\mathcal{F}^{\alpha-1}$. Kripke defines $\phi(S_1^{\alpha-1}, S_2^{\alpha-1}) = (S_1^\alpha, S_2^\alpha)$ as:

Let S_1^α be the set of (codes of) true sentences of $\mathcal{L}(S_1^{\alpha-1}, S_2^{\alpha-1})$, and let S_2^α be the set of all elements of D which either are not (codes of) sentences of $\mathcal{L}(S_1^{\alpha-1}, S_2^{\alpha-1})$ or are (codes of) false sentences of $\mathcal{L}(S_1^{\alpha-1}, S_2^{\alpha-1})$. [names of sets are altered to fit the notation in this proof]

Kripke (1975) p702

Kripke uses Kleene's strong three-valued logic to deal with connectives, standard first-order logic to deal with the semantics of ordinary predicates and defines the extension (anti-extension) of a partially defined predicate P as the set of sentences for which Px is true (false).

Hansen defines \mathcal{T}^α as $\mathcal{T}^{\alpha-1} \cup t_\alpha$ and \mathcal{F}^α as $\mathcal{F}^{\alpha-1} \cup f_\alpha$ and uses 11 rules, TE1-TE11, to define t_α and f_α . The aim is to prove that

$$\begin{aligned} s \in S_1^\alpha \setminus S_1^{\alpha-1} &\Rightarrow s \in t_\alpha \\ s \in S_2^\alpha \setminus S_2^{\alpha-1} &\Rightarrow s \in f_\alpha \end{aligned}$$

In the following, we will only consider the cases where $s \notin S_1^{\alpha-1} \cup S_2^{\alpha-1}$. By Theorem 5 it follows that if those sentences are in $S_1^\alpha \cup S_2^\alpha$, they are in $t_\alpha \cup f_\alpha$.

- If s is of the form $P(t_1 \dots t_n)$ where P is an ordinary n -ary predicate and $t_1 \dots t_n$ are closed terms, it follows from semantics for classic first order logic and TE1 that if $s \in S_1^\alpha(S_2^\alpha)$ it is also in $t_\alpha(f_\alpha)$.
- For s of the form $\neg\phi$, $s \in S_1^\alpha(S_2^\alpha)$ iff ϕ is in $S_2^\alpha(S_1^\alpha)$. If ϕ was in $S_1^{\alpha-1} \cup S_2^{\alpha-1}$ then s would have been in $S_1^{\alpha-1} \cup S_2^{\alpha-1}$, so ϕ must be in $S_1^\alpha \setminus S_1^{\alpha-1} \cup S_2^\alpha \setminus S_2^{\alpha-1}$. If $\phi \in S_1^\alpha \setminus S_1^{\alpha-1} \cup S_2^\alpha \setminus S_2^{\alpha-1}$ it also follows by induction hypothesis that $\phi \in t_\alpha \cup f_\alpha$. It follows by theorem 4 that $s \notin \mathcal{U}$, so TE2 applies to s in Hansen. Since by induction, if $\phi \in S_1^\alpha(S_2^\alpha)$ then $\phi \in \mathcal{T}^\alpha(\mathcal{F}^\alpha)$ and by TE2 it follows that if s is in $S_1^\alpha \setminus S_1^{\alpha-1}$ ($s \in S_2^\alpha \setminus S_2^{\alpha-1}$) then it is in $t_\alpha(f_\alpha)$.
- Consider the case of s of the form $(\phi \vee \psi)$. $s \in S_1^\alpha$ iff $\phi \in S_1^\alpha$ or $\psi \in S_1^\alpha$ and $s \in S_2^\alpha$ iff $\phi \in S_2^\alpha$ and $\psi \in S_2^\alpha$. By the same reasoning as for the case of s of the form $\neg\phi$, s is not in \mathcal{U} if either ϕ or ψ are not in \mathcal{U} . It follows with TE3 that if $s \in S_1^\alpha(S_2^\alpha)$ it is also in $t_\alpha(f_\alpha)$.
- The cases for the other connectives are analogous to the case for disjunction.
- Kripke states for the existential quantifier:

$(\exists x)A(x)$ is true if $A(x)$ is true for some assignment of an element of D to x ; false if $A(x)$ is false for all assignments to x , and undefined otherwise.

Kripke (1975) p700

Thus, it holds for a sentence s of the form $\exists x\phi$ that it is in S_{α_1} iff there is a constant c such that $\phi(x/c) \in S_1^\alpha$, and in S_2^α iff for all constants c , $\phi(x/c) \in S_2^\alpha$. By the same reasoning as for the case where s is of the form $\neg\phi$, $s \notin \mathcal{U}$ if there is a constant c for which $\phi(x/c) \notin \mathcal{U}$. It follows with TE7 that if $s \in S_1^\alpha(S_2^\alpha)$ it is also in $t_\alpha(f_\alpha)$.

- The case for the universal quantifier is analogous to the case for the existential quantifier.
- If s is of the form $T(t)$ and $I(t) = d$ it is in $S_1^\alpha(S_2^\alpha)$ iff d is in $S_1^\alpha(S_2^\alpha)$. As argued for s of the form $\neg\phi$, if $d \in S_1^\alpha \cup S_2^\alpha$, $s \notin \mathcal{U}$. It follows with TE9 that if $s \in S_1^\alpha(S_2^\alpha)$ it is also in $t_\alpha(f_\alpha)$.

This shows that for a successor ordinal α , language $\mathcal{L}^\alpha = \mathcal{L}(S_1^\alpha, S_2^\alpha)$ and tentative evaluation $E_{E^{\alpha-1}} = (\mathcal{T}^\alpha, \mathcal{F}^\alpha, \mathcal{U}^{\alpha-1})$, all sentences that are in S_1^α are in \mathcal{T}^α and all sentences that are in S_2^α are in \mathcal{F}^α . Since taking a tentative evaluation to an evaluation does not affect \mathcal{T} and \mathcal{F} , the relation also holds for \mathcal{L}^α and E^α .

The last case is the case of α a limit ordinal different from 0. For \mathcal{L}^α , S_1^α and S_2^α are defined as:

$$S_1^\alpha = \bigcup_{\beta < \alpha} S_1^\beta$$

$$S_2^\alpha = \bigcup_{\beta < \alpha} S_2^\beta$$

For E^α , \mathcal{T}^α , \mathcal{F}^α and \mathcal{U}^α are defined as:

$$\mathcal{T}^\alpha = \bigcup_{\beta < \alpha} \mathcal{T}^\beta$$

$$\mathcal{F}^\alpha = \bigcup_{\beta < \alpha} \mathcal{F}^\beta$$

$$\mathcal{U}^\alpha = \bigcup_{\beta < \alpha} \mathcal{U}^\beta$$

Assume for the sake of contradiction that not all sentences in $S_1^\alpha(S_2^\alpha)$ are in $\mathcal{T}^\alpha(\mathcal{F}^\alpha)$. Then there must be sentence s which is in $S_1^\alpha(S_2^\alpha)$ and not in $\mathcal{T}^\alpha(\mathcal{F}^\alpha)$. It follows that here are β' and β'' smaller than α such that $s \in S_1^{\beta'}(S_2^{\beta'})$ and $s \notin \mathcal{T}^{\beta''}(\mathcal{F}^{\beta''})$. Let β''' be the largest of these. By monotonicity, $s \in S_1^{\beta'''}(S_2^{\beta'''})$ and $s \notin \mathcal{T}^{\beta'''}(\mathcal{F}^{\beta'''})$, which contradicts the induction hypothesis. \square

This shows that *if* a proper k-sentence is in the extension (anti-extension) of the truth predicate in Kripke's minimal fixed point, then it is in the set $\mathcal{T}(\mathcal{F})$ of the Hansen total evaluation. That the implication cannot be reversed is shown

by the following set of sentences based on the strengthened Liar which are both proper K-sentences:

$$\neg T(c_{sl})$$

where $I(c_{sl}) = \neg T(c_{sl})$, and the sentence

$$T(c_1)$$

where $I(c_{nsl}) = \neg T(c_{nsl})$.

For Kripke, neither sentence is grounded because they both refer to the sentence $\neg T(c_{sl})$, which refers to itself. Because they are not grounded, they are undefined in the minimal fixed point. For Hansen, on the other hand, $\neg T(c_{sl})$ becomes undefined at level 1, but because $T(c_{nsl})$ does not depend on itself, E3 fails for $T(c_{nsl})$, so it does not become undefined at level 1 and instead becomes true at level 2.

3.2 Kripke's other fixed points

Kripke constructs a minimal fixed point as potential modelling of the concept of truth, but also discusses the possibility of constructing other fixed points by starting of the iterative process not with an empty extension and anti-extension of the truth predicate, but with some sentences already in either set at stage 0. This allows him to distinguish between grounded sentences (which have a truth value at the minimal fixed point), paradoxical sentences (which do not have a truth value at any fixed point), sentences with arbitrary truth values (which are true in some fixed points, false in others) and sentences with intrinsic truth values (which are not grounded but in every fixed point where they have a truth value, they have the same truth value).

Can the same be done in Hansen's system? That is, what happens if we start of the evaluation process with some sentences already in one of the true, false or undefined sets?

The Hansen system, as it is, behaves differently in this respect from the Kripke system. First of all, in Kripke's system, each sentence is evaluated again at each iteration. So if the Liar sentence is first put in, say, the extension of the truth predicate, then at the next level it disappears from the extension and becomes a member of the anti-extension of the truth predicate instead. This cannot happen with Hansen. With each iteration, the old sets \mathcal{T} , \mathcal{F} and \mathcal{U} are taken and then some more sentences are added to create the new sets. So if the Liar is put in the set \mathcal{T} at level 0, then at level 1, it is still there. But because the constant contained in the Liar (call it c_l) now refers to a sentence which is in the set \mathcal{T} , at TE_{E^0} the Liar sentence $F(c_l)$ *also* becomes false. So the Liar is now both true and false and the evaluation has become inconsistent.

To avoid this problem Hansen's system can be adapted so that the old sets \mathcal{T} and \mathcal{F} are discarded and only the sentences that have received a proper truth value in the tentative evaluation with respect to the old sets are part of the

new sets. This does not change the behaviour of the normal evaluation process where $E^0 = (\emptyset, \emptyset, \emptyset)$. Because $\mathcal{T}_\mathcal{E}$ and $\mathcal{F}_\mathcal{E}$ are defined as $\mathcal{T} \cup t_\mathcal{E}$ and $\mathcal{F} \cup f_\mathcal{E}$, discarding the sets \mathcal{T} and \mathcal{F} could only lead to less sentences having a proper truth value at a particular level. But for the normal evaluation process, if the truth or falsehood of a sentence could be determined at an earlier level, it still can at the present one. So the adaptation does not affect the normal evaluation process, but gives the same effect for the Liar as it has in Kripke when made either true or false: it indefinitely moves back and forth between true and false and there is no evaluation \mathcal{E} such that $E_\mathcal{E} = \mathcal{E}$ for this assignment. Sentences with arbitrary truth values and sentences with intrinsic truth values now also function like in Kripke.

In Hansen's system it is also possible to declare sentences undefined at the start, and take the iterative process from there. Doing this leads to rather awkward results: because the rules TE1-TE11 have the explicit condition that a sentence that becomes either true or false is not yet declared undefined, a sentence that starts out undefined always remains undefined. This condition cannot be removed without affecting the normal behaviour of the system: it would for instance allow the Liar to be evaluated again according to TE10 after it is already declared undefined, which would re-instate paradox.

This means that perfectly innocuous sentences like "Snow is white" can become undefined in evaluations \mathcal{E} for which $E_\mathcal{E} = \mathcal{E}$. This would also mean that "Snow is white", although grounded, does not have an intrinsic truth value. Where making ungrounded sentences true or false to see what happens leads to a natural modelling of the different types of sentences there are, allowing sentences to be ungrounded at the start does not model any intuitive properties and is better described as strange behaviour resulting from not following the rules.

Note that properties like 'arbitrary' and 'intrinsic' are meta-properties in Hansen when defined as above, just like they are in Kripke. That is, there is no way to express in the object language that the Truth Teller has an arbitrary truth value.

3.3 Semantic equivalences in Hansen and Kripke

There are many semantic equivalences for classical logic, most of which are very intuitive. Kripke's theory satisfies most equivalences between formulas, but does not satisfy all equivalences involving the notion of semantic consequence. In section 2.1.2 it was mentioned that Kripke's implication is criticised for not satisfying all intuitive criteria. This section shows that very few equivalences hold for Hansen at all. Specifically, equivalences between sentences that may or may not be self-referential do not hold in general. As shown earlier in this chapter in Section 3.1.3, for the most ordinary sentences Hansen and Kripke give the same evaluations and thus for those cases the same equivalences hold for Hansen as hold for Kripke.

The equivalences mentioned in this section are divided into propositional

equivalences, quantified equivalences, equivalences concerned with implication and logical consequence and equivalences concerning truth. For some equivalences a detailed discussion is given, others are only mentioned in a more general context. This section also provides some insight in the principles causing certain equivalences to fail and the type of sentence for which they fail.

3.3.1 Propositional equivalences

$\phi \wedge \perp \text{ Eq } \perp$

\perp is used to represent some formula that at some point becomes false.

This equivalence holds in Hansen. \perp becomes false at some level, and $\phi \wedge \perp$ becomes false at the same level by TE4. $\phi \wedge \perp$ does not become undefined because as long as $\phi \wedge \perp$ and \perp are undetermined, E3 does not hold for $\phi \wedge \perp$ because \perp is assumed to be grounded.

In Kripke, if one conjunct is false, the conjunction becomes false. So when \perp becomes false, so does $\phi \wedge \perp$ and the equivalence holds in Kripke also.

Commutativity, associativity

In Hansen, laws of commutativity and associativity for disjunction, conjunction and bi-implication do not hold. If s and s' are equivalent under commutativity and associativity, they both depend on the same set of sentences. If all those sentences depend on s , but not on s' and E1 and E4 are satisfied for both s and s' then s becomes undefined but s' may receive a proper truth value. If, for instance s is equal to $T(c) \wedge 0 = 0$ and s' is equal to $0 = 0 \wedge T(c)$ and $I(c) = T(c) \wedge 0 = 0$, s is self-referential and becomes undefined, whereas s' is not and becomes false instead.

For Kripke, because s and s' refer to the same sentences, they become undefined under the same circumstances. Also, disjunction, conjunction and bi-implication are defined so that s and s' receive the same truth value if they receive a proper truth value.

Other propositional equivalences

Many other classic equivalences do not hold in Hansen for the same reasons mentioned for commutativity and associativity. Equivalences that do not hold are for instance $\phi \vee \perp \text{ Eq } \phi$, $\phi \wedge \phi \text{ Eq } \phi$, $\neg(\phi \wedge \psi) \text{ Eq } \neg\phi \vee \neg\psi$, $\neg\phi \text{ Eq } \phi \rightarrow \perp$, distribution laws and double negation elimination.

All these equivalences do hold for Kripke because both formulas refer to the same sentences and are undefined under the same circumstances. The definitions of the connectives also guarantee that the equivalences hold for cases with proper truth values.

In the absence of self-reference

For two sentences s and s' that are both not self-referential, all classic propositional equivalences hold in Hansen. TE1-TE6 guarantee this for the cases where s and s' have subformula with proper truth values. If s and s' only still depend on sentences that are the beginning of an infinite dependency chain, E3 is satisfied for both. Since E4 is satisfied for s iff it is satisfied for s' , s and s' now become undefined under the same circumstances. This guarantees the validity of for instance $\phi \wedge \phi \text{ Eq } \phi$ in the absence of self-reference.

3.3.2 Quantified equivalences

$\forall v\phi \text{ Eq } \neg\exists v\neg\phi$

This equivalence does not hold in Hansen. An example of two formulas s and s' of the form $\forall v\phi$ and $\neg\exists v\neg\phi$ that have different truth values are the sentences $\forall x(P(x) \rightarrow U(x))$ and $\neg\exists x\neg(P(x) \rightarrow U(x))$, where P is only true of s and c_s is a constant referring to s . At TE_{E^0} , all instances of $P(x) \rightarrow U(x)$ for which x is not equal to c_s become true, and for those x the instances of $\neg(P(x) \rightarrow U(x))$ become false. At E_{E^1} , $P(c_s) \rightarrow U(c_s)$ becomes undefined and so does s but E3 fails for $\neg(P(c_s) \rightarrow U(c_s))$. At TE_{E^2} , $U(c_s)$ becomes true, so $\neg(P(c_s) \rightarrow U(c_s))$ becomes false, so $\exists x\neg(P(x) \rightarrow U(x))$ becomes false and s' becomes true. Thus, $\llbracket s \rrbracket = +$ and $\llbracket s' \rrbracket = \top$ and $\forall v\phi$ is not equivalent with $\neg\exists v\neg\phi$.

For Kripke, s is ungrounded and because s' requires a truth value for s to receive a truth value itself, s' is also ungrounded and both formulas are undefined. Since Kripke explicitly defines the universal quantifier in terms of the existential quantifier, the equivalence must hold for Kripke.

α -equivalent formulas

Some equivalences, like $\forall v\forall w\phi \text{ Eq } \forall w\forall v\phi$ and $\forall v\phi \text{ Eq } \forall wA[v : w]$ (for w not free in ϕ), depend on α -equivalence. Both equivalences hold in Hansen if α -equivalent formulas are referred to by the same constant, but otherwise one formula may be undefined and the other have a proper truth value. It is not clear whether Hansen intended this kind of identity, for a discussion on this topic, see section 2.3.2.

Because in Kripke, sentences making claims about undefined sentences are themselves undefined, the identity issue does not arise and both sentences are undefined under the same circumstances. Also, when they have a proper truth value, they have the same proper truth value.

Other quantified equivalences

Because self-reference may succeed with one formula but fail with the other, other quantified equivalences also fail to hold in general. Mentioned here are $\forall v\phi \text{ Eq } \phi$ (where v is not free in ϕ) and $\forall v(\phi \wedge \psi) \text{ Eq } (\forall v\phi \wedge \forall v\psi)$. These equivalences do hold for Kripke.

In the absence of self-reference

TE7 and TE8 together with TE2 and TE4 show that if all instances of a quantified formula have a proper truth value, the equivalences in this paragraph hold. Also, if a particular amount of instances have received a proper truth value and this is enough to give the sentence on one side of the equation a proper truth value, this is also the case for the other side, provided it has not yet become undefined. Also, if an infinite dependency chain originating from one of the instances of the formula on one side of the equation causes it to be undefined, this will also be the case for the other side. Thus, in the absence of self-reference, the equivalences do hold in Hansen.

3.3.3 Implication

$\phi \models \psi$ Eq $\models \phi \rightarrow \psi$

\Rightarrow The equivalence does not hold in this direction in Hansen. Let ϕ be the sentence $\neg T(c)$ and ψ be the formalisation of ‘Snow is green’, $\forall x(S(x) \rightarrow G(x))$, where $I(c) = \neg T(c) \rightarrow \forall x(S(x) \rightarrow G(x))$. Assume ψ to be false. $\neg T(c) \rightarrow \forall x(S(x) \rightarrow G(x))$ now directly depends only on $\neg T(c)$, which only depends on sentences again depending on $\neg T(c) \rightarrow \forall x(S(x) \rightarrow G(x))$. ϕ is now undefined in all models, so $\phi \models \psi$ is true, but $\neg T(c) \rightarrow \forall x(S(x) \rightarrow G(x))$ is also undefined in all models and $\models \phi \rightarrow \psi$ is not true.

\Leftarrow This does hold in Hansen. If $\phi \rightarrow \psi$ is true in all models, then in all models either ϕ is false or undefined, or ψ is true. In both cases it is true that if $M \models \phi$ then $M \models \psi$, so $\phi \models \psi$ is true.

For Kripke, if ϕ is ungrounded and ψ is false, then $\phi \rightarrow \psi$ is ungrounded, so for such ϕ and ψ , $\phi \models \psi$ does hold, but $\models \phi \rightarrow \psi$ does not. However, if $\phi \rightarrow \psi$ is valid, then by semantics of implication $\phi \models \psi$ is also true.

$\models \phi \rightarrow \phi$

This does not hold in Hansen. If ϕ is equal to $\neg T(c)$ and $I(c) = \neg T(c) \rightarrow \neg T(c)$, then $\phi \rightarrow \phi$ is equal to the interpretation of c , and because all formulas it depends on depend on the interpretation of c , $\phi \rightarrow \phi$ always depends on itself and becomes undefined. Therefore, $\phi \rightarrow \phi$ is not true for all ϕ and all models.

For Kripke, if ϕ is ungrounded then so is $\phi \rightarrow \phi$ and $\models \phi \rightarrow \phi$ does not hold.

$\phi, \phi \rightarrow \psi \models \psi$

This does hold in Hansen. If both ϕ and $\phi \rightarrow \psi$ are true in a model, then by TE5, ψ also has to be true, and $\phi, \phi \rightarrow \psi \models \psi$ holds. The same is true for Kripke and his definition of implication.

3.3.4 Truth

$$T(\ulcorner \phi \wedge \psi \urcorner) \text{ Eq } T(\ulcorner \phi \urcorner) \wedge T(\ulcorner \psi \urcorner)$$

This equivalence holds in Hansen. Note that because quotation names are used, $T(\ulcorner \phi \wedge \psi \urcorner)$, $T(\ulcorner \phi \urcorner)$ and $T(\ulcorner \psi \urcorner)$ all have proper truth values (see section 2.4.3). If $\phi \wedge \psi$ is undefined, then so are both ϕ and ψ and the equivalence holds. If $\phi \wedge \psi$ is false, then one of ϕ and ψ is false or undefined, call this formula χ . Then $T(\ulcorner \chi \urcorner)$ is false and $T(\ulcorner \phi \urcorner) \wedge T(\ulcorner \psi \urcorner)$ is false. If $\phi \wedge \psi$ is true, then both ϕ and ψ must be true and the equivalence holds.

Because the equivalence holds for all possible truth values of $T(\ulcorner \phi \wedge \psi \urcorner)$, it holds in general.

These sentences cannot be formulated in Kripke's system, as it lacks a quotation device.

$$T(c_{\phi \wedge \psi}) \text{ Eq } T(c_\phi) \wedge T(c_\psi)$$

This equivalence must be understood with $I(c_{\phi \wedge \psi}) = \phi \wedge \psi$ for arbitrary ϕ and ψ , $I(c_\phi) = \phi$ and $I(c_\psi) = \psi$. It does not hold in Hansen. If both ϕ and ψ are equal to $T(c)$, and $I(c) = T(c_{\phi \wedge \psi})$, then $T(c_{\phi \wedge \psi})$ only depends on itself and it is undefined, and the same holds for $T(c)$, but this makes $T(c_\phi)$ and $T(c_\psi)$ both false, and thus $T(c_\phi) \wedge T(c_\psi)$ becomes false.

For Kripke, both sentences are undefined under the same circumstances, and have the same proper truth value if they have proper truth values. Thus, the equivalence holds in Kripke.

Names and equivalences

The example of the two equivalences mentioned in this paragraph, which are identical except for different kinds of names for formulas, highlights some relevant issues for equivalences. First of all, the failed equivalence using constant names shows the same structure as the failing propositional and quantified equivalences: one formula may be self-referential, while the other is not. This principle also causes other equivalences using the truth predicate to fail.

Second of all, the difference between the equivalences show that this issue does not arise for quotation names. Self-referential formulas where the self-reference occurs within the quotation marks, in the absence of infinite dependency chains, receive a proper truth value and this causes the equivalence to hold. This is a principle that can be generalised to propositional and quantified equivalences: if one of the sentences is self-referential, but the self-reference occurs within quotation marks, the equivalence still holds. So almost all equivalences fail for all-names sentences, but hold for quotation-names sentences.

3.4 Conclusion

In this chapter we have considered the relationship between Hansen's recent proposal and Kripke's earlier work that inspired Hansen. It was shown that all

sentences that are grounded according to Kripke and therefore receive a proper truth value in his system, are also grounded according to Hansen and receive the same proper truth value. This is a desirable property of both systems: the sentences that are grounded according to Kripke are intuitively unproblematic and should have a proper truth value, as they indeed have. In Hansen's system however, some more sentences receive proper truth values because he employs a wider notion of groundedness that allows sentences to be grounded in the ungroundedness of other sentences.

Hansen's system turns out to allow for the construction of alternative fixed points like those Kripke uses to classify sentences as grounded, intrinsic, paradoxical and as having an arbitrary truth value. This construction does require a minor technical adaptation which does not affect the ordinary 'minimal fixed point' process of the system. Starting off with sentences already declared undefined does not yield any intuitive results, and it is not possible to alter the theory to change these results without having the whole system completely collapse. Thus, it is possible to do the same in Hansen's system as is possible in Kripke's system, but the extra set of undefined sentences does not give an advantage here.

Hansen's implication fails in the same way Kripke's implication fails in the sense that it does not satisfy $\models \phi \rightarrow \phi$. In both cases it can be argued that, given pathological examples like the Liar, this is not completely counter-intuitive. Hansen fails to satisfy almost all classical equivalences between formulas, which do hold for Kripke. This is due to reference occurring outside quotation marks where a constant or predicate may be defined so that it refers to one sentence but not the other. This causes one sentence to be self-referential, while the other is not. The one sentence will now become undefined while the other only refers to an undefined formula or has an undefined subformula and receives a proper truth value based on that. For Kripke, because both formulas refer to undefined formulas, both are undefined. These results are further evaluated in Section 4.2.

Chapter 4

A critical view

“The model then, is to be tested by its technical fertility. It need not capture every intuition, but it is hoped that it will capture many.” - Kripke (1975)

In the previous sections some topics have been touched upon that may impact the value ascribed to the system. This chapter reflects on the technical value of the behaviour of the system with regard to certain types of formulas as described in section 2.4 and critically evaluates the outcomes for the equivalences in section 3.3. It will look at the valuation schemes used by Hansen and investigates whether or not Strong Kleene would be a reasonable alternative. A critical note to the intuitiveness of the evaluation process is placed in section 4.5. The next section will look at the failing of the Leibniz law in Hansen’s system and how this can be fixed. Section 4.7 looks at the definition of quantifiers and the possibility of using domain constants. The section 4.9 considers criterion E4 and proposes an adaptation. The last section mentions a shortcoming of the theory already noted by Hansen concerning generalisations over the semantics of the system.

4.1 Undefined formulas

Section 3.1.3 shows that all K-grounded sentences have a proper truth value in Hansen, which means that the system works as one would hope for the set of indisputably ordinary cases. Section 2.4 shows that also sentences where the pathology (i.e. self-reference or infinite dependency) occurs only inside quotation marks receive a proper truth value. The extent to which this is a desirable property will be discussed in section 4.9 below.

Section 2.4 also shows that a formula containing undefined subformulas may itself have a proper truth value. It seems odd that, for instance, “The Liar is false” and “The Truth Teller is true” are both undefined, but “The Liar is false and the Truth Teller is true” is simply false. Surely, a conjunction of

two pathological cases is itself pathological? This property also makes having a proper truth value non compositional in some sense: a sentence may have a proper truth value, while its subformula do not. The status of these sentences is further discussed in section 4.4.

Sentences that fail to have a proper truth value themselves may have subformula that do have proper truth values. This makes sense, for instance the Disjunctive Liar, “This sentence is false or $0=1$ ”, is clearly pathological, whereas its part “ $0=1$ ” is ordinary. Subformulas that do not have a truth value yet at the moment a sentence becomes undefined become undefined at the same time. This does seem intuitive: if a sentence had undetermined subformulas that were ever going to have proper truth values the sentence should not yet become undefined because possibly those subformulas would make the sentence true or false. By contraposition, if a sentence does become undefined, it makes sense that its undetermined subformula must also be given up on at that time.

Sentences that closely resemble undefined sentences may themselves have proper truth values, as the case of the formula $\neg T(c_{nsl})$ and $\neg T(c_{sl})$ with $I(c_{nsl}) = I(c_{sl}) = \neg T(c_{sl})$ shows. This may be confusing at first, but there is an actual ambiguity to Liar-type sentences, since for instance the Strengthened Liar does have a reading for which it is true, namely the reading where it is considered as a sentence claiming of some untrue sentence (namely, the Strengthened Liar) that it is not true. This ambiguity is neatly reflected in $\neg T(c_{nsl})$ and $\neg T(c_{sl})$, where the latter is the Strengthened Liar and the former truthfully says of the Strengthened Liar that it is not true. So, the sentence $\neg T(c_{nsl})$, although closely related to a sentence with problematic self-reference, is itself quite ordinary in judging the truth value of some other sentence, and justly receives the ordinary truth value true. The modelling of this reading and the semantic behaviour of this formula are expanded upon in the next section.

4.2 Semantic equivalences

Section 3.3 presents some classic or intuitive equivalences and discusses whether or not they hold in Hansen. As it turns out, equivalences between formulas that are not guaranteed to have a proper truth value never hold. This is due to the fact that reference to a formula is something very specific: a name always refers to a single formula. This means that for any two intuitively equivalent formulas, one may be self-referential while the other is not. It is for this reason that most equivalences fail: as shown in section 3.1.3, sentences that receive a proper truth value in Kripke receive the same truth value in Hansen. Thus, for ordinary cases, most equivalences hold like they do in Kripke. The equivalences also hold for sentences containing only quotation names, as they always receive a proper truth value. Also, sentences that are undefined due to infinite dependency chains satisfy intuitive equivalences because their equivalents are also undefined for the same reasons. Thus, only cases where one formula is self-referential while the other is not cause the equivalences to fail.

But, one might say, these equivalences do hold for Kripke, so why not do

in Hansen whatever it is that makes it work for Kripke? The reason some equivalences hold in Kripke that do not hold in Hansen, is that for Kripke, sentences referring to undefined sentences are themselves undefined, whereas in Hansen, these sentences may have a proper truth value. This allows Hansen to, for instance, declare the fact that the Strengthened Liar is not true using $\neg T(c_{nsl})$. The fact that Hansen can express the fact that the Strengthened Liar is not true is an advantage. But the formula $\neg T(c_{nsl})$ is not the only way to express this fact, Hansen's theory naturally provides an alternative: $\neg T(\ulcorner T(c_{sl}) \urcorner)$. E4 guarantees that this sentence receives a proper truth value and TE2 and TE9 make it true. Using the quotation device also gives a natural expression to the difference between $\neg T(c_{sl})$, which *is* the Strengthened Liar, and $\neg T(c_{nsl})$ and $\neg T(\ulcorner T(c_{sl}) \urcorner)$, which attempt to speak *about* the Strengthened Liar. The external perspective the latter two sentences take with respect to the Strengthened Liar is intuitively expressed by the use of quotation. Still, it seems to be possible in natural language to express the fact that the Strengthened Liar is not true simply using the sentence “The Strengthened Liar is not true”, and one is not committed to saying “‘The Strengthened Liar is not true’ is not true”. So, in natural language we can express the the Strengthened Liar not being true using a sentence similar to $\neg T(c_{nsl})$ and it also seems reasonable to give this sentence a proper truth value. This happens in the current definition of Hansen's system, but is also the reason some intuitive equivalences fail and part of the reason Leibniz's law fails (see section 4.6).

4.3 Choice of negation

Kripke uses choice negation, which means that the negation of an undefined sentence is itself undefined. Hansen, on the other hand, uses exclusion negation, where the negation of an undefined sentence is true.

First of all, Kripke does not really have much choice on the matter. That is, when a sentence is ungrounded by Kripke's definitions, so is its negation.¹ Hansen does have a choice: the liar sentence $F(c_l)$ with $I(c_l) = F(c_l)$ is undefined because it depends on itself, but the negation of the Liar, $\neg F(c_l)$ depends on the Liar, but not on itself. Because the Liar is undefined, the negation of the Liar can be either true or undefined (or false, but that is not a very intuitive option). Hansen chooses to make this sentence true, arguing that

$$\text{It is not the case that the Liar is false} \tag{4.1}$$

is intuitively true because the Liar is undefined. There are at least three ways to achieve this result. The first is the way Hansen himself seems to choose: formalising (4.1) as $\neg F(c_l)$ and using exclusion negation. The second way is to formalise (4.1) as $\neg T(\ulcorner F(c_l) \urcorner)$ and using any negation, interpreting “it is not the

¹This is true in the minimal fixed point. Using choice negation instead of exclusion negation protects Kripke from having all negated sentences receiving proper truth values at the first stage because their negated formulas are still undefined.

case that” as “it is not true that”. $F(c_l)$ now becomes undefined, $T(‘F(c_l)’)$ false and $\neg T(‘F(c_l)’)$ true. This is defensible for the case of (4.1), but there is also a sense in which

$$\text{The Liar is not false} \tag{4.2}$$

is true, and then the second formalisation becomes rather counter-intuitive. This leads to the third option: formalising (4.1) (or (4.2)) as $\neg F(c_{nl})$ where $I(c_{nl}) = F(c_l)$, and using any negation. Because $F(c_l)$ is undefined, $F(c_{nl})$ is false and $\neg F(c_{nl})$ is true. The fact that $\neg F(c_l)$ is only true when the negation symbol is interpreted as exclusion negation and $\neg F(c_{nl})$ is true either way points to a flaw in the first formalisation. This is best explained using the sentence

$$\text{The Liar is false.} \tag{4.3}$$

There is a sense in which this sentence is false: since the Liar is undefined, it is not false. But we cannot use $F(c_l)$ to express this fact because since $I(c_l) = F(c_l)$, this sentence would itself be the liar sentence. If we want to claim the falsehood of the Liar, we have to use another constant whose interpretation is $F(c_l)$, for instance c_{nl} and claim that $F(c_{nl})$. In the first formalisation of (4.2), the supposed falsehood of the Liar is expressed exactly by the inappropriate sentence $F(c_l)$ and the exclusion negation is used as a cheap fix to still obtain the desired truth value. The third formalisation of (4.2) shows our desire to speak *about* the Liar sentence already in its formalisation of the sentence that is negated, “The Liar is false”. This formalisation does not force a choice between exclusion and choice negation, since $F(c_{nl})$ already has a proper truth value.

This shows that when properly formalised, sentence (4.1) does not require exclusion negation to become true. That does not mean that exclusion negation is a bad choice: it simply shows that it would need another justification to be anything other than arbitrary.

4.4 Strong Kleene

Considering that choice negation must, at least for the moment, be considered a reasonable alternative to exclusion negation, is it also reasonable to combine Strong Kleene valuations with Hansen’s theory?

For Strong Kleene valuations, besides choice negation, disjunction is defined so that a disjunction is true if one of its disjuncts is true, false if both are false and undefined otherwise. To extend this to Hansen, a policy for undetermined is required, and Hansen’s ‘undefined’ is identified with Kleene’s ‘undefined’². The

²The way Kripke uses Kleene and undefinedness, undefined means both that a sentence does not yet have a truth value (during the construction) and that it will never receive one (at the fixed point). This distinction is made explicit in Hansen by using undetermined for the ‘not yet’ interpretation and undefined for the ‘not ever’ interpretation. Thus, the identification of Kleene undefined with Hansen undefined, given Kripke, is not trivial.

most natural option for undetermined is that a disjunction is undetermined if it has an undetermined disjunct whose future truth value is relevant for the truth value of the disjunction. The truth tables for conjunction, implication and equivalence can be defined in the usual way with the truth tables of negation and disjunction.³

In the previous section it was argued that exclusion negation may not be preferable over choice negation since a proper modelling of natural language utterances results in equivalent results for both. This extends to the definitions of the other connectives: it is generally a sign of bad modelling when an undefined sentence turns up as subformula in a larger sentence that is not itself intended to be pathological. Pathological sentences should become undefined according to E1-E4 and non-pathological sentences should not have pathological subformula. For instance, consider the case of the disjunction of the Liar and the Truth Teller:

$$\text{“The Liar is false or the Truth Teller is true”}. \quad (4.4)$$

One may want to formalise this as:

$$F(c_l) \vee T(c_t)$$

with $I(c_l) = F(c_l)$ and $I(c_t) = T(c_t)$. Both disjuncts become undefined at the first level. According to Hansen’s truth tables, the disjunction now becomes false, but according to Strong Kleene, it becomes undefined. In order to establish which truth value is more natural, let’s first consider what actually happens in this formula. Because both disjuncts have the self-reference typical for the Liar and the Truth Teller, they do not speak *about* those sentences, but in a sense, they actually *are* the Liar and the Truth Teller. Translated back into natural language, rather than (4.4), it looks most like:

$$\text{“This disjunct is false or this disjunct is true”}.$$

Given the clear self-reference in both disjuncts, it seems rather odd to declare this sentence simply false: surely, it is more complicated than that, and declaring the sentence undefined is more intuitive. But what about (4.4)? There certainly is a sense in which the sentence is false because of the reading where the first disjunct reflects on the Liar and (falsely) claims that it is false, and the second reflects on the Truth Teller and (again, falsely) claims that it is true. This reading is formalised by:

$$F(c_{nl}) \vee T(c_{nt})$$

³As Hansen points out in a footnote, even though the truth tables of the connectives can be derived from each other, the corresponding equivalence between formulas does not hold, that is, $\phi \rightarrow \psi$ and $\neg\phi \vee \psi$ are not equivalent for arbitrary ϕ and ψ . This is evident from the failed equivalences in section 3.3.

with $I(c_{nt}) = F(c_t)$ and $I(c_{nt}) = T(c_t)$, and is indeed false, simply because both disjuncts are false. Note that this sentence is false in Hansen, regardless of whether one employs Hansen's own valuation scheme or the adapted Strong Kleene, but undefined in Kripke anyway. There is an actual advantage Hansen has over Kripke, but it is simply not situated in the valuation scheme used.

The example above argues for using an adapted version of Strong Kleene instead of the scheme Hansen suggests. Despite the intuitive result this gives for this example, it rather messes up the architecture of the system. Each level originally was divided in two parts: the tentative evaluation and the evaluation. At the tentative evaluation step, sentences were added to the sets \mathcal{T} and \mathcal{F} , and only those sets: no sentences became undefined at that step. Then, an evaluation was made where sentence were added to \mathcal{U} , and \mathcal{T} and \mathcal{F} were left alone. This clear separation of when and how sentences receive proper truth values and when and how they become undefined disappears if Strong Kleene is introduced, since sentences can then become undefined at the tentative evaluation in addition to the evaluation.

4.5 Stability of truth values

In Kripke's system the Liar, $\neg T(c_{sl})$, with $I(c_{sl}) = \neg T(c_{sl})$, is ungrounded and therefore undefined. If we were, against all regulations, to re-evaluate the Liar given that knowledge, that is, if we were to evaluate $\neg T(c_{sl})$ knowing that c_{sl} refers to some undefined sentence, $\neg T(c_{sl})$ would remain undefined, because $T(c_{sl})$ and therefore also $\neg T(c_{sl})$ would only end up in the extension or anti-extension of T if the sentence c_{sl} refers to were in either, but it was determined that c_{sl} referred to an undefined sentence. The truth values of sentences are consistent in that sense for Kripke.

This is not so for Hansen. The Strengthened Liar (in Hansen's system the Strengthened Liar and the Liar are different formulas, but the same point could be made for both) is undefined, because the constant c_{sl} refers to a sentence $\neg T(c_{sl})$ whose truth-value cannot be determined prior to giving a truth-value to the Strengthened Liar, $\neg T(c_{sl})$. But if we were to illegally re-evaluate $\neg T(c_{sl})$ after it has been determined that c_{sl} refers to an undefined sentence, it would turn up true. So the truth value of sentences is not always stable in this sense. The claim is not that there is actual inconsistency in Hansen's system or that the Strengthened Liar is actually true, but there is an intuitive consistency in truth values in Kripke's system that Hansen's system lacks. For Kripke, once you know the truth value of a sentence, it continues to make sense, but for Hansen, one has to stop thinking about the truth value of a sentence once it is determined or it may cease to make sense.

One could argue that this is not a flaw, because there actually is a reading for the Strengthened Liar that makes it true: the Strengthened Liar is in fact not true (for it is undefined) and this is what is reflected in the new truth value of the Strengthened Liar if it gets illegally re-evaluated. The proper way to express this fact in Hansen is to use another formula, for instance $\neg T(c_{nsl})$,

which is not identical to the Strengthened Liar but says of the Strengthened Liar that it is not true. That this is possible is a strength of Hansen's system, but it does not come without the intuitive instability of truth values and the failing of Leibniz law, and that just does not seem quite right. The Leibniz law will be the subject of the next section.

4.6 The Leibniz law

The Leibniz Law does not hold in Hansen. Consider the Strengthened Liar, $\neg T(c_{sl})$, where $I(c_{sl}) = \neg T(c_{sl})$. Next, consider the quoted sentence $\ulcorner \neg T(c_{sl}) \urcorner$. It now holds that $c_{sl} = \ulcorner \neg T(c_{sl}) \urcorner$, but not that $\neg T(c_{sl}) \leftrightarrow \neg T(\ulcorner \neg T(c_{sl}) \urcorner)$ since $\neg T(c_{sl})$ is undefined and $\neg T(\ulcorner \neg T(c_{sl}) \urcorner)$ is true.

Hansen argues that this principle could be restored by stipulating that when a sentence becomes undefined, so do all sentences which are identical with it except for different terms with the same reference⁴. In the example of the Strengthened Liar, this would mean that when $\neg T(c_{sl})$ becomes undefined, so does $\neg T(\ulcorner \neg T(c_{sl}) \urcorner)$, and $\neg T(c_{sl}) \leftrightarrow \neg T(\ulcorner \neg T(c_{sl}) \urcorner)$ holds.

This, Hansen argues, comes at the cost of some expressive strength: it is now not possible anymore to express the intuitive fact that the Strengthened Liar is actually not true, even though it remains possible to express the fact that it is undefined.

But this opens up for another possibility to express that the Strengthened Liar is not true. Hansen proves that in the unique total evaluation, all sentences are either true, false or undefined.⁵ This means if the Strengthened Liar is not true, it is either false or undefined. In Hansen's adapted language, the claim that the Strengthened Liar is not true, $\neg T(\ulcorner \neg T(c_{sl}) \urcorner)$, turns up undefined, but the claim that it is either false or undefined, $F(\ulcorner \neg T(c_{sl}) \urcorner) \vee U(\ulcorner \neg T(c_{sl}) \urcorner)$, which is of an entirely different form than the Strengthened Liar, becomes true because $U(\ulcorner \neg T(c_{sl}) \urcorner)$ does.

This indicates that Hansen's adapted language can in fact express the same facts as the original Hansen language, but in a seemingly less intuitive way. That is, it is not possible anymore to express the fact that the Strengthened Liar is not true directly anymore: it first needs to be translated to the fact that the Strengthened Liar is either undefined or false. This may seem like a loss, but it is in fact not very different from Hansen's original construction where the fact that the Strengthened Liar is not true cannot simply be expressed with the sentence $\neg T(c_{sl})$, but can be expressed with for instance $\neg T(c_{nsl})$. It is essential for any sentence expressing something about an undefined sentence, that it is not identical to the undefined sentence itself, or it would be undefined. The adapted Hansen language uses the same principle, but it stretches the definition of 'identical' a bit, in a not completely counter-intuitive way. Now, the only way to express the Strengthened Liar not being true without creating a sentence that

⁴Hansen (2014), p235

⁵Hansen (2014), p229

is identical to the Strengthened Liar itself, is to claim that the Strengthened Liar is either false or undefined.

4.7 Domain constants

If a model is not guaranteed to have constant names for all elements in the domain, domain constants are usually added to govern the behaviour of quantifiers. Quantification in Hansen is defined using the constants specified by the model. This leads to the behaviour quantifiers usually have because all elements in the domain are assigned at least one constant. Is this equivalent to an approach using domain constants, that is: a set of constants that covers the domain but may be a subset of the set of all constants, \mathcal{C} ? The answer is no.

Consider the Strengthened Liar, $\neg T(c_{sl})$, and two constants referring to the Strengthened Liar, c_{sl} and c_{nsl} . Additionally, consider the quantified sentence:

$$\forall x(P(x) \rightarrow \neg T(x)), \quad (4.5)$$

where P is a predicate only true of $\neg T(c_{sl})$.

If all constants are considered as in Hansen's definition, then (4.5) depends on $P(c_{sl}) \rightarrow \neg T(c_{sl})$. Because $P(c_{sl})$ is true and $\neg T(c_{sl})$ is undefined, this instance is false and (4.5) is also false.

With domain constants, if of the two only c_{nsl} is a domain constant, (4.5) becomes true. The instance of (4.5) for c_{nsl} is $P(c_{nsl}) \rightarrow \neg T(c_{nsl})$. Because c_{nsl} refers to an undefined formula, $\neg T(c_{nsl})$ is true and the whole formula is true. Because $P(x)$ only holds for c_{nsl} , all other instances of (4.5) are also true and (4.5) becomes true.

This shows that the approach with domain constants differs from the approach taken in Hansen. Hansen's approach, given the functioning of the system, can be argued to be better by giving a more complete view of all possible instances of a formula. On the other hand, using the external and internal distinction made in section 4.4, that is, the distinction between creating a self-referential sentence and referring to a self-referential sentence, one could argue that any instance of a quantified formula should always be the external version of a sentence, not the internal one.

The effect of letting the instances of quantified formulas only be external versions of sentences is easily achieved by introducing a new set of domain constants. This set precisely covers the domain, that is: each element in the domain has exactly one corresponding domain constant. The set of domain constants is stipulated to be disjoint from the set of ordinary constants so that no domain constant refers to a sentence that depends on a sentence containing that domain constant, because there are no such sentences. Domain constants are then only used to define quantifiers, guaranteeing an external perspective. Again, whether or not this is desirable can be debated.

4.8 Satisfaction

Although theories of truth are more popular, theories of satisfaction have some advantages. First of all, satisfaction can be defined on open formulas in addition to sentences, whereas truth can only be defined on sentences. Secondly, satisfaction can be made compositional in a sense truth most certainly cannot be. This section will investigate the possibility of creating a satisfaction predicate in a fashion similar to the way Hansen defines truth and compare the two systems. First, the usual form of the satisfaction predicate is introduced. Next, a system similar to Hansen's system for truth is defined for satisfaction. Finally, the new system for satisfaction is compared to the original system for truth.

4.8.1 The satisfaction predicate

Where truth is a one-place predicate defined on sentences, satisfaction is a two-place predicate defined on assignment functions and formulas. The satisfaction predicate typically looks something like:

$$Sat(\alpha, \phi)$$

where α is an assignment function and ϕ a formula possibly containing free variables. Several design choices can be made for the assignment function, all having pro's and con's, and we will work with an assignment function that is defined precisely on the free variables of ϕ . We can now use $Sat([x : d], P(x))$ to express that $P(x)$ is true of object d ⁶. Systems using satisfaction require of the domain not only that it contains all sentences of the language, but also all open formulas and all variables. These elements are also required to have names.

Truth corresponds to a satisfaction relation between a sentence and the empty assignment. Because truth is defined on sentences it makes sense to have it correspond to satisfaction of sentences. Because sentences do not have any free variables and assignments were stipulated to be defined precisely on the free variables of a formula, the assignment corresponding to a sentence is the empty assignment, which will be denoted with ε . Thus, a theory of satisfaction corresponds to a theory of truth if it satisfies

$$T(\bar{s}) \Leftrightarrow Sat(\varepsilon, s')$$

for all sentences s , where \bar{s} is a name for s and s' the equivalent of s in a language using satisfaction. It will turn out in section 4.8.3 that a theory of satisfaction based on Hansen's theory of truth does not satisfy this equality.

4.8.2 Hansen-satisfaction

This section aim to define satisfaction in a way similar to the way Hansen defines truth. It will look at the major changes in Hansen's system this entails. It is

⁶The assignments are assignments from variables to domain objects. In this formula d is an actual domain object, not a constant name.

not in itself a complete theory, additional material from Hansen's definition that does not require adaptation is still necessary for a complete description.

The predicates

Just like Hansen employs a falsity and undefinedness predicate alongside a truth predicate, we now also require two predicates in addition to the satisfaction predicate, to express that a certain assignment makes a formula false or undefined. These predicates will be named *FOf* and *UOf*, to express formulas being false of and undefined of certain objects. For the sake of consistency, *Sat* is renamed as *TOf*. We now get:

$$\begin{aligned} &TOf(\alpha, \phi) \\ &FOf(\alpha, \phi) \\ &UOf(\alpha, \phi) \end{aligned}$$

for our truth, falsity and undefinedness predicates respectively.

Concepts

A number of concepts defined in Hansen for truth need to be redefined for satisfaction. First of all, quotations disappear. Satisfaction does not make use of constant names but variables and assignments to domain objects instead.

Because an open formula does not have a truth value in itself, in the next section evaluations will be made on tuples of formulas and assignments. Since the concept of dependency was used to aid the evaluation process, in particular, the part of the process where sentences are declared undefined, it will be used in the new system to aid the part of the process where tuples of formulas and assignments become undefined. It therefore makes sense to define dependency on these tuples. The new definition of dependency runs as follows:

The binary relation $R_{\mathcal{E}}$ on $\mathcal{S} \times \mathcal{A}$, called the *direct dependency relation with respect to the evaluation \mathcal{E}* , is defined as follows: $(\phi, \alpha)R(\psi, \beta)$ if both (ϕ, α) and (ψ, β) are undetermined according to \mathcal{E} , and

- ϕ is $\neg A$, ψ is A and α is β ,
- ϕ is $(A \vee B)$, $(A \wedge B)$, $(A \rightarrow B)$ or $A \leftrightarrow B$, ψ is A or B and $\alpha \upharpoonright FV(\psi) = \beta$,
- ϕ is $\exists xA$ or $\forall xA$, ψ is A and $\beta \upharpoonright FV(\phi) = \alpha$, or
- ϕ is $TOf(\gamma, A)$, $FOf(\gamma, A)$ or $UOf(\gamma, A)$ and ψ is A .

Let $\overline{R}_{\mathcal{E}}$, the *dependency relation with respect to the evaluation \mathcal{E}* , be the transitive closure of $R_{\mathcal{E}}$.

Finally, as mentioned in the previous paragraph, the *domain*, D , is a superset of the set of formulas and the set of variables. The set of constants includes names for all elements of the domain.

Evaluations

Because open formula do not have truth values by themselves, formulas are evaluated together with an assignment function. The sets $t_{\mathcal{E}}$ and $f_{\mathcal{E}}$ are not filled by requiring certain formula are elements of those sets but by requiring that certain sets of tuples of formulas and assignments are subsets of those sets. In the following new definitions of TE1-TE11, $dom(f)$ is a function that returns the domain of a function f and $f \upharpoonright A$ gives the function identical to the function f but with a domain limited to the set A .

TE1) If ϕ is of the form $P(t_1, \dots, t_n)$ where P is an ordinary n-ary predicate and t_1, \dots, t_n are terms, then:

- $\{(\phi, \alpha) \mid \alpha : FV(\phi) \rightarrow d, (\llbracket t_1 \rrbracket_{\alpha}, \dots, \llbracket t_n \rrbracket_{\alpha}) \in I(P)\} \subseteq t_{\mathcal{E}}$
- $\{(\phi, \alpha) \mid \alpha : FV(\phi) \rightarrow d, (\llbracket t_1 \rrbracket_{\alpha}, \dots, \llbracket t_n \rrbracket_{\alpha}) \notin I(P)\} \subseteq f_{\mathcal{E}}$

TE2) If ϕ is of the form $\neg\psi$ where $\phi \notin \mathcal{U}$ then

- $\{(\phi, \alpha) \mid \alpha : FV(\phi) \rightarrow d, (\psi, \alpha) \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}\} \subseteq t_{\mathcal{E}}$
- $\{(\phi, \alpha) \mid \alpha : FV(\phi) \rightarrow d, (\psi, \alpha) \in \mathcal{T}_{\mathcal{E}}\} \subseteq f_{\mathcal{E}}$

TE3) If ϕ is of the form $(\psi \vee \chi)$ where $\phi \notin \mathcal{U}$ then

- $\{(\phi, \alpha) \mid \alpha : FV(\phi) \rightarrow d, (\psi, \alpha \upharpoonright FV(\psi)) \in \mathcal{T}_{\mathcal{E}}\} \cup \{(\phi, \alpha) \mid \alpha : FV(\phi) \rightarrow d, (\chi, \alpha \upharpoonright FV(\chi)) \in \mathcal{T}_{\mathcal{E}}\} \subseteq t_{\mathcal{E}}$
- $\{(\phi, \alpha) \mid \alpha : FV(\phi) \rightarrow d, (\psi, \alpha \upharpoonright FV(\psi)) \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}, (\chi, \alpha \upharpoonright FV(\chi)) \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}\} \subseteq f_{\mathcal{E}}$

TE4) $(\psi \wedge \chi)$ follows from TE2 and TE3.

TE5) $(\psi \rightarrow \chi)$ follows from TE2 and TE3.

TE6) $(\psi \leftrightarrow \chi)$ follows from TE2 and TE3.

TE7) If ϕ is of the form $\exists v\psi$ where $\phi \notin \mathcal{U}$ then:

- $\{(\phi, \alpha) \mid \alpha : FV(\psi) \setminus \{x\} \rightarrow d, \exists \beta : FV(\psi) \rightarrow d, \beta \upharpoonright dom(\alpha) = \alpha, (\psi, \beta) \in \mathcal{T}_{\mathcal{E}}\} \subseteq t_{\mathcal{E}}$
- $\{(\phi, \alpha) \mid \alpha : FV(\psi) \setminus \{x\} \rightarrow d, \forall \beta : FV(\psi) \rightarrow d, \beta \upharpoonright dom(\alpha) = \alpha, (\psi, \beta) \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}\} \subseteq f_{\mathcal{E}}$

TE8) $\forall v\psi$ follows from TE7.

TE9) If ϕ is of the form $TOf(\alpha, \psi)$ where $\phi \notin \mathcal{U}$ then:

- $(\phi, \varepsilon) \in t_{\mathcal{E}}$ if $(\psi, \alpha) \in \mathcal{T}$
- $(\phi, \varepsilon) \in f_{\mathcal{E}}$ if $(\psi, \alpha) \in \mathcal{F} \cup \mathcal{U}$

TE10) If ϕ is of the form $FOf(\alpha, \psi)$ where $\phi \notin \mathcal{U}$ then:

- $(\phi, \varepsilon) \in t_{\mathcal{E}}$ if $(\psi, \alpha) \in \mathcal{F}$
- $(\phi, \varepsilon) \in f_{\mathcal{E}}$ if $(\psi, \alpha) \in \mathcal{T} \cup \mathcal{U}$

TE11) If ϕ is of the form $UOf(\alpha, \psi)$ where $\phi \notin \mathcal{U}$ then:

- $(\phi, \varepsilon) \in t_{\mathcal{E}}$ if $(\psi, \alpha) \in \mathcal{U}$
- $(\phi, \varepsilon) \in f_{\mathcal{E}}$ if $(\psi, \alpha) \in \mathcal{T} \cup \mathcal{F}$

The construct the evaluation $E_{\mathcal{E}}$ relative to the evaluation \mathcal{E} , tuples (ϕ, α) satisfying all of the following are added to the set \mathcal{U} :

- E1) $(\phi, \alpha) \notin \mathcal{T}_{\mathcal{E}} \cup \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}$
 E2) $TE_{\mathcal{E}}|\overline{R}_{\mathcal{E}}(\phi, \alpha) = \mathcal{E}|\overline{R}_{\mathcal{E}}(\phi, \alpha)$,
 E3) for every tuple (ψ, β) , if $(\phi, \alpha)\overline{R}_{\mathcal{E}}(\psi, \beta)$ then $((\psi, \beta)\overline{R}_{\mathcal{E}}(\phi, \alpha)$ or there is an infinite $\overline{R}_{\mathcal{E}}$ -sequence $(\psi, \beta)\overline{R}_{\mathcal{E}}(\chi, \gamma)\overline{R}_{\mathcal{E}} \dots$ consisting of distinct elements), and

As mentioned, quotation names disappear when employing satisfaction, and so E4 disappears with them.

Truth values

The process described in the previous section creates sets \mathcal{T} , \mathcal{F} and \mathcal{U} of tuples of formulas and assignments. Such a tuple can be called true, false or undefined if it features in the corresponding set and undetermined if it features in none of them.

The interpretation of a formula can now be given based on the assignments that make it true, false or undefined respectively. Define the interpretation of a formula ϕ at a level α as:

$$\llbracket \phi \rrbracket^{\alpha} = (\mathsf{T}^{\alpha}, \mathsf{F}^{\alpha}, \mathsf{U}^{\alpha})$$

where $\mathsf{T}^{\alpha} = \{\alpha \mid (\phi, \alpha) \in \mathcal{T}^{\alpha}\}$ and analogously for F^{α} and U^{α} .

Using the definition above, we can speak of sentences being true, false, undefined or undetermined if their interpretation corresponds to $(\{\varepsilon\}, \emptyset, \emptyset)$, $(\emptyset, \{\varepsilon\}, \emptyset)$, $(\emptyset, \emptyset, \{\varepsilon\})$ or $(\emptyset, \emptyset, \emptyset)$ respectively. It is easy to see that a sentence only has one assignment function, ε , which features in at most one of the sets T , F and U .

4.8.3 Names

Names play an important role in Hansen's theory: for a formula $F(\overline{F(c_l)})$ it is rather essential which name of $F(c_l)$ is used. If one uses the name c_l , the formula becomes undefined. If one uses the name c_{nl} or ' $F(c_l)$ ', it becomes false. For the sentence $F(\overline{T(c)})$ where c refers to the formula itself, using a constant name for $T(c)$ makes the sentence undefined, whereas using quotation names makes it false.

This distinction cannot be made using satisfaction. Satisfaction does not use names to assign objects to argument positions in formulas, but assignments from variables to objects instead. Hansen argues that it makes sense for a formula

to be evaluated only after all formula it quotes are evaluated, but this principle cannot be replicated with satisfaction in any straightforward way. More importantly, the difference between for instance the Strengthened Liar ($\neg T(c_{sl})$) and the sentence claiming of the Strengthened Liar that it is not true ($\neg T(c_{nsl})$) does not have an equivalent for satisfaction. This effect is quite similar to what happens if the Leibniz law is restored for Hansen's system with truth, as is discussed in section 4.6. A piece of expressive strength that is retained is that it is still possible to say of the Liar sentence⁷ that it is undefined. If l is the Liar sentence, then $UOf(\varepsilon, l)$ is the sentence saying of the Liar that it is undefined.

Because the distinction between self-referential and other constant names is no longer made, the issue raised in section 4.7 about domain constants no longer applies to the semantics of quantifiers.

These considerations imply that although it is possible to create a satisfaction predicate in a way similar to Hansen's construction of a truth predicate, the two give different results for sentences.

4.8.4 Compositionality

One of the advantages satisfaction has over truth is that satisfaction can be compositional where truth is not. This has to do with the definition of quantifiers. Compositionality for semantics entails that the meaning of composite expression can be determined based on the meaning of its parts and the way they are composed. In a definition of truth, like the one Hansen gives, this does not hold for quantified formulas. For instance, the truth value of $\forall xP(x)$ directly depends on $P(c)$, even though $P(c)$ is not a part of $\forall xP(x)$. Satisfaction on the other hand is compositional for this example in the sense that the semantics of $\forall xP(x)$ can be determined based on the semantics of $P(x)$.

However, using satisfaction does not make Hansen's semantics compositional. This is because there is another reason besides the semantics of quantifiers that makes Hansen's theory non-compositional, namely the process by which sentences become undefined. Complex sentences do not become undefined as a function of the truth values of their composite parts (the truth tables in section 2.2.2 show that this never happens), but based on information about the sentences that these parts depend on. Using satisfaction instead of truth does not change this fact.

4.9 E4

E4 is the fourth criterion a sentence has to satisfy for Hansen in order to become undefined. Where E1 is a technical condition guaranteeing the stability of truth

⁷The Liar as presented in this thesis does not exist in the new satisfaction language, since it employs a falsity predicate is not part of the new language. One way to construct paradox using satisfaction is by mimicking Grellings heterological paradox. The Grelling Paradox and its formalisation are discussed in Visser (1989), p621.

values, E2 is included in E3 and E3 mirrors the forms of all known Liar paradoxes, E4 is intended to capture an intuitive idea that is specific for Hansen's theory. This section critically evaluates the results it gives for different types of sentences and suggests an alternative version of the criterion.

4.9.1 E4 and circularity

Criterion E4 stipulates that a sentence s only becomes undefined if there is no sentence s' which is quoted in s such that $s\overline{R}s'$. Hansen argues the desirability of this criterion with the sets of sentences

$$(4.7) \text{ is false} \tag{4.6}$$

$$(4.6) \text{ is true} \tag{4.7}$$

and

$$(4.9) \text{ is false} \tag{4.8}$$

$$\text{"(4.9) is false"} \text{ is true.} \tag{4.9}$$

He argues that in the case of (4.9) it is intuitive to evaluate the quoted sentence (4.8) prior to evaluating the whole of (4.9) in a way that is not there for (4.7) and (4.6). Thus, (4.8) becomes undefined and (4.9) becomes false.

There seems to be some intuitive truth in the idea that when a sentence quotes another sentence, the quoted sentence should be evaluated before the quoting sentence. However, it is not very intuitive that something sensible can be said about the truth value of a sentence prior to that sentence receiving a truth value. That is, it is not very intuitive that "(4.9) is false" can be undetermined without giving up on (4.9), but actually assigning it a proper truth value at a later stage. Thus, whether or not E4 gives intuitive results for circular sentences is at least debatable and, I would argue, the results are simply not intuitive.

4.9.2 E4 and infinity

Necessary...

All Yablo sentences are undefined and it would be a feature if the system could express this fact. Hansen shows that this is actually possible, using the sentence:

$$\forall y(N(y) \rightarrow U(\neg\forall P(y, x) \rightarrow \neg T(x))) \tag{4.10}$$

where $I(N) = \mathbb{N}$, $I(P) = \{(n, Ym) \mid n, m \in \mathbb{N} \text{ and } m > n\}$ and Y_i the i 'th Yablo sentence. The evaluation of the Yablo sentences is discussed in section 2.4. At E_{E^1} , (4.10) only still depends on all Yablo sentences, and since all

Yablo sentences are the beginning of an infinite dependency chain consisting of distinct elements, E3 is satisfied for (4.10). The only thing stopping the sentence from becoming undefined is E4, which forces the sentence to be evaluated after all Yablo sentences have received a truth value.

For any sentence purporting to say something about the Yablo sequence without being part of it itself, there needs to be some way of lifting the sentence above the Yablo sentences and it seems natural enough to use the quotation device to do this. Also, the intuitive paradoxicality of sentence (4.9) does not appear here. Because none of the Yablo sentences depends on sentence (4.10), their truth value can be determined independently of (4.10).

...but not sufficient

In Hansen's system, (4.10) is true, but this cannot be expressed in the system using constant names. That is,

$$T(c_{4.10}) \tag{4.11}$$

where $c_{4.10}$ denotes (4.10), is undefined. Why is this? (4.11) directly depends only on (4.10), but because of that it also indirectly depends on all sentences (4.10) depends on, that is: all Yablo sentences and everything they depend on. By the usual Yablo-reasoning, (4.11) does not become undefined at E_{E^0} because E2 fails. At E_{E^1} , (4.11) depends only still on (4.10) and all Yablo sentences. E3 is satisfied for (4.11) as all Yablo-sentences are the beginning of an infinite dependency chain and so is (4.10). E4 is also satisfied for (4.11). Since E1 is satisfied for (4.11) as well, it becomes undefined.

If it is possible to express the truth of (4.10), but that would require using quotes, so $T(\ulcorner \forall y(N(y) \rightarrow U(\ulcorner \forall P(y, x) \rightarrow \neg T(x) \urcorner)) \urcorner)$ is true. Thus, if one wants to speak about (4.11), one has to continue using quotes to not fall back into the infinite hole of the Yablo paradox.

4.9.3 A new E4

The examples of sentences (4.9) and (4.10) show that E4 can have both intuitive and counter-intuitive effects, depending on whether the quoting sentence depends on itself through the quoted sentence. To accomplish intuitive results in both cases, re-define E4 as E4':

E4') If there is a sentence s' which is quoted in s such that $s\overline{R}s'$ then $s'\overline{R}s$.

This invalidates Hansen's theorem 9.2 concerning the intermediate Tarskian quotation names T -, F - and U -schemata. In particular, the formalisation of the sentence

“‘This sentence is undefined’ is undefined”

is undefined, even though the “This sentence is undefined”-part is actually undefined. This outcome relies on the intuition that it is not reasonable to evaluate claims about the truth value of a sentence before evaluating the sentence itself and that sentences that attempt to force this should be undefined.

Sentence (4.11) calls for an additional adaptation of E4 to include sentences that depend on quoting sentences. Formerly, this would have been problematic because sentences with self-reference within quotation marks would never receive a truth value as their quoted sentence would not, but since circularity is not covered anymore in E4, no such problems occur and all sentences still receive a truth value at some point⁸. E4 now becomes:

E4'') If there are sentences s' and s'' such that $s\overline{R}s'$ and s'' is quoted in s' and $s\overline{R}s''$, then $s''\overline{R}s$.

This validates a stronger version of the external all-names schemata than were previously valid. Hansen proves for his own theory:

Theorem 6. *For every model $\mathfrak{M} = (D, I)$, sentence s and constant c such that $I(c) = s$, the following holds:*

- If $\llbracket T(c) \rrbracket = \top$, then $\llbracket s \rrbracket = \top$. If $\llbracket s \rrbracket = \top$, then $\llbracket T(c) \rrbracket \in \{\top, +\}$.
- If $\llbracket F(c) \rrbracket = \top$, then $\llbracket s \rrbracket = \perp$. If $\llbracket s \rrbracket = \perp$, then $\llbracket F(c) \rrbracket \in \{\top, +\}$.
- If $\llbracket U(c) \rrbracket = \top$, then $\llbracket s \rrbracket = +$. If $\llbracket s \rrbracket = +$, then $\llbracket U(c) \rrbracket \in \{\top, +\}$.

With E4'' it holds that:

Theorem 7. *For every model $\mathfrak{M} = (D, I)$, sentence s and constant c such that $I(c) = s$, the following holds:*

- If $\llbracket T(c) \rrbracket = +$, then $\llbracket s \rrbracket = +$
- If $\llbracket F(c) \rrbracket = +$, then $\llbracket s \rrbracket = +$
- If $\llbracket U(c) \rrbracket = +$, then $\llbracket s \rrbracket = +$

Proof. Let X stand for any of T , F and U . Assume at absurdum that s receives a proper truth value at some level $\alpha + 2$. Then at $\alpha + 1$, it did not satisfy all of E1-E4''. The following shows that if one of E1-E4'' was not satisfied for s , it was not for $X(c)$ either.

- If E1 was not satisfied for s , s was an element of $\mathcal{T}_{\alpha+1} \cup \mathcal{F}_{\alpha+1} \cup \mathcal{U}$, which contradicts our assumption that s receives a proper truth value at level $\alpha + 2$. Therefore, E1 must have held for s .
- Theorem 3 shows that the case for E2 follows from the case for E3.

⁸A formal proof of this property will not be given here, but it is intuitively true that the new system retains it.

- This holds by theorem 1.
- If $E4''$ failed for s then there must have been sentences s' and s'' such that $s\overline{R}_{E^\alpha}s'$ and s'' is quoted in s' and $s\overline{R}_{E^\alpha}s''$, but not $s''\overline{R}_{E^\alpha}s$. Because $(X(c))\overline{R}_{E^\alpha}s$, also $(X(c))\overline{R}_{E^\alpha}s'$, so if, ad absurdum, $E4''$ held for $X(c)$ then $s''\overline{R}_{E^\alpha}(X(c))$ must have held. But then also $s''\overline{R}_{E^\alpha}s$, which contradicts our assumptions. Therefore, if $E4''$ failed for s , it also failed for $X(c)$.

Because not all of $E1$ - $E4''$ were satisfied for s at level $\alpha + 1$ they were not for $X(c)$ either and by $E1$, $X(c)$ is undetermined at the beginning of level $\alpha + 2$. Because s receives a proper truth value at $\alpha + 2$, $X(c)$ receives a proper truth value at $\alpha + 3$.

This shows that if s receives a proper truth value, so does $X(c)$, so by contraposition, if $X(c)$ does not receive a proper truth value, then neither does s . \square

From this follows that:

Theorem 8. *For every model $\mathfrak{M} = (D, I)$, sentence s and constant c such that $I(c) = s$, the following holds:*

- $\llbracket T(c) \rrbracket = \top$, iff $\llbracket s \rrbracket = \top$
- $\llbracket F(c) \rrbracket = \top$ iff $\llbracket s \rrbracket = \perp$
- If $\llbracket U(c) \rrbracket = \top$, then $\llbracket s \rrbracket = +$. If $\llbracket s \rrbracket = +$, then $\llbracket U(c) \rrbracket \in \{\top, +\}$.

Proof. For the first two bullets it holds that if the left hand side holds, then also the right hand side by theorem 6. If the right hand side holds, combining theorem 6 and 7 gives the left hand side. The last bullet comes directly from theorem 6. \square

The bullet for $U(c)$ cannot be strengthened because s does not have a proper truth value and so theorem 7 does not contradict the possibility of $U(c)$ being undefined.

Comparing $E4''$ to $E4$, $E4''$ does not allow conclusions being drawn about the truth value of a sentence before evaluating the sentence itself, whereas $E4$ does. $E4''$ also contrasts with $E4$ in making it possible to sensibly speak about sentences that speak about sentences with infinite dependency chains without continuing to use quotes. $E4''$ invalidates a version of the T-schema that $E4$ does validate, but $E4''$ also validates two versions of the T-schema that $E4$ does not validate.

4.10 Intuitively true generalisations

The previous section mentioned some different versions of Tarski's schemata Hansen distinguishes between. The full range of versions Hansen employs consists of eighteen versions: six for each semantic predicate, divided up into three

versions for sentences only containing quotation names and three versions for sentences with all names. These three versions distinguish between different degrees of use of a meta-language, resulting in internal, intermediate and external versions.

All all-names schemata fail. Hansen argues that this is in fact reasonable, using philosophical justification to argue that classical logic, unrestricted validity of the T-schemata and semantic closure cannot hold jointly. For quotation names, the external and intermediate T-schemata hold, but the internal quotation names schemata fail. Hansen describes the internal quotation names schema and its behaviour:

The internal quotation names T-schema can be formulated as $\forall v(P(v) \rightarrow T(v))$, where P is a unary predicate such that $I(P)$ is the set of all sentences of the form $s \leftrightarrow T(\ulcorner s \urcorner)$ where s is a sentence. In the present theory, this sentence becomes undefined. The problem is that $\forall v(P(v) \rightarrow T(v))$ can only become true after all instances of $s \leftrightarrow T(\ulcorner s \urcorner)$ have been made true. And one of these instances is the one where s is $\forall v(P(v) \rightarrow T(v))$. The same holds *mutatis mutandis* for the internal F - and U -schemata.

Hansen (2014), p238

Hansen continues to admit that this is a genuine shortcoming of his theory:

More generally, this theory shares the problem with Kripke's theory that intuitively true generalisations about the whole semantics are not made true by the theory.

Hansen (2014), p238

These “intuitively true generalisations” do not only include the internal T-schemata, but also for instance the intuitively true sentence mentioned in Gupta (1982): “No sentence is both true and false”. That this is a property of Hansen's system follows from his lemma 5.2 where all evaluations are proven to be consistent. It can be formalised in the object language as $\neg\exists v(S(v) \rightarrow (T(v) \wedge F(v)))$, where S is a predicate true only of all sentences. Because $T(v) \wedge F(v)$ is false for all other sentences, the truth of this sentences essentially depends on itself and it becomes undefined, not true.

4.11 Conclusion

In this chapter a wide variety of topics has been discussed. Section 4.1 showed that many tricky sentences receive intuitive truth values. Some modelling issues of sentences in natural language that clarified the relationship between formulas and natural language and thus their intuitive meaning were mentioned in sections 4.2, 4.3 and 4.4. Section 4.2 reflected on the failure of many semantic equivalences and showed that this is intrinsically linked with features of the system like the expressibility of the lack of truth for the Strengthened Liar. Sections 4.3 and 4.4 looked at the possible alternative of using Strong Kleene valuations

in Hansen's system. It turned out that this would lead to some more intuitive results, but decrease the elegance of the system. Section 4.5 suggested that the evaluations in the system are unstable in a sense but that this, again, is linked to the expressive strength of the system. Section 4.6 showed that the Leibniz law, which fails in Hansen's system, can be restored by making more sentences undefined, demanding that some natural language sentences are modelled in a less intuitive way. Section 4.7 showed that using domain constants in the semantics of quantifiers is not equivalent with the definition used in Hansen and could both be argued to be inferior and to be superior. Section 4.9 showed that E4 has some more and less intuitive results and suggested an alternative which gives more intuitive results and validates different versions of the T-schemata. The last section re-states a shortcoming already mentioned by Hansen, namely that some intuitively true generalisations about the whole semantics are not made true by the theory.

Chapter 5

Conclusion

“When I use a word,’ Humpty Dumpty said in rather a scornful tone, ‘it means just what I choose it to mean - neither more nor less.’” - Lewis Carroll (1871)

After a substantial amount of words have been spent discussing the merits of Hansen’s theory, this chapter aims to give some guidance in processing everything that has been written here. The first section gives a summary of the previous chapters, only mentioning the topics discussed and without going into any detail. The next section reflects on the thesis itself and the process of its creation and its significance. Section 5.3 provides some suggestions for further research, while section 5.4 discusses the consequences for the field of Artificial Intelligence. The last section answers the research question in the most general terms.

5.1 Summary

In the past 4 chapters the theory of truth presented by Hansen has been thoroughly analysed in its definitions and results, accomplishments have been discussed, criticism has been given and motivated and adaptations have been suggested.

5.1.1 Hansen’s theory

A general discussion of the concepts Hansen defines has been offered in chapter 2. This chapter also shows the results the theory gives for certain interesting sentences. The degree to which these results are desirable is up for discussion and has been discussed in section 4.1 and various other sections that are referenced in section 4.1. Most notable is the sentence claiming of the Strengthened Liar that it is not true, which is undefined for Kripke but true for Hansen. That Hansen can express the fact that the Strengthened Liar is not true is a major

accomplishment, but it is paid for by a certain instability of truth values (section 4.5), the failing of almost all classical semantic equivalences (section 3.3) and the failing of Leibniz law (section 4.6). The result the theory gives for sentences consisting of undefined sentences like the conjunction of the Liar and the Truth Teller, which is false in Hansen's theory, was questioned in section 4.4.

Chapter 2 also discusses criterion E2 Hansen devises for deciding whether or not a sentence should become undefined at a particular level. Section 2.5 shows that E2 is formally redundant because it is included in E3. However, E2 is not unpleasant to use when manually evaluating sentences, because if it fails (and by implication, E3 fails), this is easier to see than the fact that E3 fails.

5.1.2 Hansen and Kripke

Chapter 3 compares Hansen's theory with Kripke's. It shows that Hansen's theory shares an important feature with Kripke's theory: the sentences that are grounded in the way Kripke defines groundedness, sentences of which it is intuitive to assume they have a proper truth value, are actually given a proper truth value in Hansen. Also, the possibility there is in Kripke's system to create other fixed points with different properties can be replicated for Hansen. As mentioned earlier, as a price for additional expressive strength, Hansen's theory sacrifices the validity of a large number of semantic equivalences that do hold for Kripke and does not satisfy any equivalences that Kripke's theory does not satisfy. As discussed in sections 3.3 and 4.2, this failure is due to a small class of sentences in Hansen's language, namely the self-referential ones employing constant names, whereas the equivalences do hold for straightforward sentences like Kripke's grounded sentences but also more pathological cases involving infinity and quotation names.

5.1.3 A critical view

The 4th chapter discusses a number of properties of Hansen's system, focussing on the controversial and undesirable. Many of the topics discussed there have already been mentioned due to their close relationship with topics raised in earlier chapters. Besides an evaluation of the results shown earlier for certain undefined formulas and semantic equivalences, the chapter also questions the valuation scheme used by Hansen, discusses an irksome counter-intuitiveness in the behaviour of truth values assigned by Hansen's system, discusses the failing of the Leibniz law and the suggestion Hansen makes to repair this, considers the possibility of using domain constants to define the semantics of quantifiers, criticises the current definition of E4 and suggests an alternative and mentions an objection already raised by Hansen: some intuitively true sentences are not made true by the theory. A recurring theme in all these sections is the relationship between natural language and certain formulas, that is, the question what certain formulas actually express and consequently, what their intuitive truth value is. Several versions of Tarski's convention T are referenced in sections 4.9 and 4.10.

5.2 About this thesis

This section reflects on this thesis from a meta-perspective. To be more precise and avoid potential problems involving self-reference, it will reflect only on the previous chapters of this thesis.

5.2.1 Method

The insights that have led to the previous chapters of this thesis are the results of careful studying of, first of all, Hansen (2014) but also Kripke (1975) and the several other works mentioned in the bibliography. These other works, together with Albert Visser's extensive knowledge of the field, have been the source of many questions that can be asked about any theory of truth, some of which proved to have interesting answers. They also served as a source of inspiration for sentences whose modelling may be counter-intuitive. Simply evaluating a lot of those sentences and further sentences inspired by the particularities of the system has led to some interesting more general insights and questions about the results of the theory and the way it is defined. Formal proof methods were employed to confirm and more than once dis-confirm certain hypotheses.

5.2.2 Objective

A theory of truth can be judged based any number of criteria. Doubtlessly, other issues can be raised and topics mentioned here can be further investigated. There is no such thing as an exhaustive investigation into the merits of a system, and this thesis does not purport to give any for Hansen's system. It is merely an attempt to discuss the most relevant issues and to provide the reader with some idea as to how the system must be evaluated. It is my belief that this thesis does in fact give an accurate overview of the main strengths and weaknesses of the system proposed. An evaluation based on this overview in general terms will be provided in section 5.5.

5.3 Further research

As mentioned in the previous paragraph, an investigation into the merits of a system is never completed. This section gives some suggestions of topics that may deserve more attention.

5.3.1 Philosophy

This thesis has mainly focussed on the technical aspects of the theory and how it relates to our logical intuitions concerning certain sentences. It has mostly left alone philosophical considerations regarding the structure of the world in general and semantics in particular. Hansen provides a metaphor to create a basis for the technical and philosophical idea of his system and the choices made there. This metaphor has been recounted in section 2.2.1 but not further

discussed in this thesis. The questions raised in primarily chapter 4 suggest that there is not always a clear answer to a question based on technical grounds and intuitions concerning specific sentences and philosophical arguments may shed some additional light on the matter. Also, the philosophical implications of the theory for concepts like that of proposition or truth itself could prove to be interesting.

5.3.2 Effects of adaptations

A number of possible adaptations of the theory have been proposed in this thesis, some of which come with a higher recommendation than others. In the relevant sections, the implications of these adaptations are touched upon, but a thorough investigation of their effects on the totality of utterances and the definition of the system is likely to turn up some additional insight.

Also, the adaptations suggested have been suggested separate from each other: no well-defined alternative theory that best reflects the preferences of the author is given. It is possible that combining multiple suggestions leads to unexpected results. Interested researchers are encouraged to pick their favourites among the alternatives given here and study the resulting theory.

5.3.3 Applications

Section 1.4, where the relationship between this thesis and the field of Artificial Intelligence was discussed, mentioned some related problems like the paradox of the Knower. This thesis focussed solely on the merits of the system with regard to defining truth, but it could be interesting to look at its applications to other problems, like the paradox of the Knower or the halting problem, or any of the other paradoxes in formal languages.

5.3.4 Satisfaction

The treatment of satisfaction in section 4.8 is limited. It gives a good basis for defining satisfaction similar to the way Hansen defines truth and some of the differences between the two, but a more thorough investigation into the system as defined in 4.8 may reveal surprising properties and possible improvements.

5.4 Impact on Artificial Intelligence

This thesis is not in itself very constructive: it does not propose any spectacular new theories or fill any gaps for theories of truth. It is probably best considered a review that got slightly out of hand. Therefore, it does not affect Artificial Intelligence in any obvious way. The relevance of the studies of truth for AI has been discussed in some depth in section 1.4 and included possible applications to other paradoxes, bridging the gap between humans and computers and modelling the world. This thesis contributes to AI by contributing to the studies of

truth: it provides a reading guide for Hansen (2014) and highlights problems that can arise for theories of truth. It can serve as reference material for those who attempt to solve one of the most basic problems in human and artificial understanding: what is truth?

5.5 How well does Hansen's theory model truth and avoid paradox?

Whether or not there is such a thing as truth, the idea of truth or third realm where truth exists is a contingent matter. If there is such a thing as truth, it is doubtful whether our use of the word 'true' is consistent with this idea, and if it is, it may or may not be possible to give a formal and general description of that use. If there is such a thing as 'the definition of truth', this is not it, and given the way the flaws and the strengths of the theory are interwoven, it may not be the way to get there either. But if there is no such perfect definition, well, then this seems a pretty good alternative.

Bibliography

- [1] JC Beal, editor. *Revenge of the Liar*. Oxford University Press, 2007.
- [2] JC Beall and Michael Glanzberg. Liar paradox. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014.
- [3] Thomas Bolander. Self-reference. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition, 2014.
- [4] Hartry Field. *Saving Truth from Paradox*. Oxford University Press, 2008.
- [5] Anil Gupta. Truth and Paradox. *Journal of Philosophical Logic*, 11(1):1–60, 1982.
- [6] Anil Gupta and Nuel Belnap. *The Revision Theory of Truth*. The MIT Press, 1993.
- [7] Casper S. Hansen. Grounded Ungroundedness. *Inquiry*, 57(2):216–243, 2014.
- [8] Hans G. Herzberger. Naive Semantics. *The Journal of Philosophy*, 11(1):61–102, 1982.
- [9] Saul Kripke. Outline of a Theory of Truth. *The Journal of Philosophy*, 72(19):690–716, 1975.
- [10] Henry Prakken. Commonsense Reasoning and Argumentation, 2014.
- [11] Alfred Tarski. Der wahrheitsbegriff in den formalisierten sprachen. *Studia Philosophica*, 1:261–405, 1935. Originally published in Polish in 1933.
- [12] Albert Visser. Semantics and the Liar Paradox. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, volume 4, pages 617–706. D. Reidel Publishing Company, 1989.
- [13] Stephen Yablo. Paradox without Self-reference. *Analysis*, 53(4):251–252, 1993.