J Philos Logic (2015) 44:681–695
DOI 10.1007/s10992-015-9353-y

One Hundred Years of Semantic Paradox

Leon Horsten

Received: 16 January 2013 / Accepted: 15 June 2013 / Published online: 26 March 2015

Abstract This article contains an overview of the main problems, themes and theories relating to the semantic paradoxes in the twentieth century. From this historical overview I tentatively draw some lessons about the way in which the field may evolve in the next decade.

Keywords Truth · Semantic paradox · Liar paradox

1 Introduction

Philosophers have been exercised by the liar paradox and its relatives since antiquity. In this article, I will not review the history of the semantic paradoxes from its beginnings. Instead, I will give an overview of the theories that describe the semantic paradoxes in the framework of modern mathematical logic. So the discussion will be restricted to the history of the semantic paradoxes in the twentieth century. Truth theory is one of the oldest and most mature disciplines of modern philosophical logic. It serves as a paradigm of the way in which a logical perspective can shed light on a fundamental philosophical concept.

I will review how new ideas and techniques of mathematical logic were put to work in the theory of truth. The focus will be on the main theories in the field. From each family of approaches, I will choose a representative, rather than discuss all or even most theories belonging to the family. There will be no space to enter into the details of the theories that are discussed in this article.

A version of this article was presented in the CAPE *Truth and Logic* Workshop at the University of Kyoto (February, 2013). I am grateful to the audience for helpful comments and suggestions.

L. Horsten (✉)
University of Bristol, Bristol, UK
e-mail: Leon.Horsten@bristol.ac.uk

In the interest of saving space (and because we will not be occupied by technical details of theories), I will be sloppy with coding. We will be concerned with proof systems and models for the 'language of truth' $\mathcal{L}_T$, which consists of the language of arithmetic (to encode syntactical notions) plus a primitive truth predicate T.

In the final section, I will tentatively develop a perspective on the future of theories of truth. This is of course a risky undertaking, for history may prove me wrong.

I will conclude this introduction with a few key bibliographical references. [7] is a superb historical account of the relation between the set theoretic paradoxes and the semantical paradoxes in the twentieth century. The collection [26] contains a bibliography of books and articles in philosophical logic on subjects related to the semantical paradoxes in the twentieth century until 1970. For an excellent systematic exposition of contemporary semantical theories about the semantic paradoxes, see [41]. On the proof theoretic side, the reader is referred to [16].

2 Before Tarski

Philosophers have constructed theories of truth since antiquity. But none of these can be classified as *formal* theories of truth in the modern sense of the word. Before the twentieth century, theories of truth were firmly located in philosophy, mostly within metaphysics or epistemology. As is typical of philosophical theories, they were not spelled out with mathematical precision.

The connection between truth and logic was not well understood. It was not generally expected of a theory of truth, for instance, systematically to explain how the truth value of a (simple or complex) sentence depends on the semantic values of its parts. This is due to various factors. First, there was no general agreement about the category to which truth belongs: is it an object, a property (of what?), a relation (between what kinds of entities?)? Some regarded truth as a special kind of entity (Frege); others regarded truth as a relation between a proposition and a fact (Aquinas). Second, logic was of course not mathematised before the middle of the nineteenth century. This made the relation between truth and logic much harder to see than it is now.

In the hands of Frege and others, logic was transformed in the nineteenth century. They expressed the laws of logic in mathematical terms, and moved logic as a discipline closer to mathematics. From around the turn of the twentieth century onwards, the field of theories of truth slowly migrated closer to mathematical logic and thereby closer to mathematics.

In the foundations of mathematics, paradoxes have played a different role than they have often played in philosophy. According Kant, for instance, certain paradoxical forms of reasoning (the antinomies of reason) ought to be accepted and learned from rather than to be dissolved. In mathematical circles, it had long been recognised that the mathematical study of paradoxes was important to make conceptual progress on foundational issues. But there was a more robust optimism that paradoxes were there to be solved in precise mathematical terms, and that by solving them, fundamental conceptual progress can typically be made. Zeno's paradoxes of motion, for instance, were taken to be resolved by the discovery of elementary principles of real

analysis, and paradoxes within the foundations of the newly discovered analysis were in turn taken to be resolved by the modern concept of limit.

During the decades of the so-called foundational crisis in mathematics, developments in the foundations of mathematics were closely connected to the study of a class of paradoxes. The paradoxes in question are related to logic and set theory, but at that time a clear separation between logic and set theory was not yet made. Old paradoxes were re-evaluated, and new paradoxes were discovered. Cantor challenged the accepted diagnoses of the known paradoxes of infinity, and Burali-Forti and Russell discovered new ones. Richard formulated a new paradox about definability, Grelling's paradox was discovered, and Russell formulated a semantic paradox about propositions that was similar to the paradox about the set of all non-self-membered sets. The argumentation of the good old liar paradox was recognised to be very similar to the pattern of reasoning in some of the new paradoxes.[1]

Progress on the set theoretic paradoxes was then made. Russell developed his theory of types to block the argument of the set of all non-self-membered sets. Zermelo and others formulated the axioms of Zermelo-Fraenkel set theory with Choice (ZFC). The upshot was that in the resulting theories—that were described with mathematical precision—the problematic paradoxical sets could not be shown to exist. This was due to either syntactic restrictions on formula construction (Russell), or to the restriction of informal comprehension principles (Zermelo).

These developments took place in the first decades of the twentieth century. At the end of it, it was fairly generally accepted that the mathematical concept of set is essentially paradox-free. The Russell paradox and the Burali-Forti paradox were taken to be solved. Some other paradoxes were not resolved. But it gradually became clear that they essentially involve extra-mathematical notions. Richard's paradox, for instance, involves the notion of definability. This notion in turn involves the notion of satisfaction, which is a generalisation of the notion of truth ('true of'). But foundational theories (such as ZFC) can be precisely articulated without making use of the notion of truth or satisfaction. Eventually, this led Ramsey to clearly distinguish between what he called the *logical paradoxes* and the *semantical paradoxes* [34]. The logical (or better: set theoretic) paradoxes were taken to be resolved, whereas the foundations of mathematics could afford to ignore the semantical paradoxes (and did so). There was a feeling (not universally shared) that because of the similarity of the structure of the arguments, the semantical paradoxes ought to be resolved along similar lines as the so-called logical paradoxes. But no solutions to the semantic paradoxes were obtained that met with widespread approval.

3 Tarski

Tarski did formulate a theory of truth that he (and most other analytical philosophers) regarded as an important contribution to philosophy. But he was first and foremost

[1] For a detailed discussion of the role of these paradoxes during the period 1900–1930, see [7, Section 2].

a mathematical logician who was obsessed with precision. He did not settle for less than a theory of truth that was spelled out in detail and with mathematical precision.

In his landmark [36], Tarski treated truth as a *property* of sentences of a language.[2] Since then, it has become fairly generally accepted in the field of formal truth theories that truth is a property of entities that are structured roughly in the way that sentences are syntactically structured.

Next, Tarski formulated a fundamental constraint that any satisfactory theory of truth for a language $\mathcal{L}$ must meet. This is his *material adequacy condition*, also known as *Convention T*.[3] It consists of all sentences of the form

p is a true sentence of $\mathcal{L}$ if and only if ϕ,

where p is a name for ϕ,[4] and the schematic variable ϕ ranges over all sentences of $\mathcal{L}$.

Tarski went on to explain how a *definition* of the concept of being a true sentence of $\mathcal{L}$ can be expressed in a specific language $\mathcal{L}^*$. The language $\mathcal{L}^*$ is taken to be an *interpreted* language that comes with a deductive apparatus S (a formal theory). Tarski then showed how all instances of Convention T can be derived from a *consistent* theory S when truth is defined in $\mathcal{L}^*$ as explained in his article.

Tarski also *proved* in the same article that no sufficiently strong consistent theory S can prove all instances of Convention T for the whole language $\mathcal{L}$ in which S is formulated. This is *Tarski's theorem on the undefinability of truth*. We cannot consistently have a materially adequate definition of the notion of true $\mathcal{L}$-sentence *in* $\mathcal{L}$ itself. It entails that the language $\mathcal{L}^*$ in which the notion of being a true sentence of $\mathcal{L}$ is formulated (the *metalanguage*) must be 'stronger' than the language $\mathcal{L}$ for which the notion of truth is defined (the *object language*). In other words, an adequate truth definition for the metalanguage $\mathcal{L}^*$ would have to be formulated in an even stronger meta-meta-language, and so on, ad infinitum. This diagnosis of and solution to the liar paradox is of course reminiscent of Russell's type-theoretic diagnosis of the paradox of the set of all non-self-belonging sets ('no definition can include the defined object in the range of its quantifiers').

Tarski did not claim that his theory of truth faithfully captures the notion of truth as used in ordinary language. The reason is that in ordinary language—let us call this $\mathcal{L}_O$—no systematic distinction is drawn between the object language and the metalanguage. Tarski claimed that in our canons of reasoning S_O of $\mathcal{L}_O$ we tend to accept even the instance of

p is a true sentence of $\mathcal{L}_O$ if and only if ϕ

where ϕ is the liar sentence (and p a name for it), for the resulting biconditional also belongs to $\mathcal{L}_O$. The argument of the liar paradox then shows that S_O must be inconsistent. Tarski's definition of truth is therefore best understood as a *rational reconstruction* in the sense of Carnap of our ordinary conception of truth.

[2] To be more precise, he took the property 'is true' to be defined in terms of the more basic satisfaction relation ('true of').

[3] I ignore Tarski's *formal* adequacy condition here, which is less important.

[4] To be more precise, p is required to be a *structurally descriptive* name.

4 In the Wake of Tarski

The turn in mathematical logic from proof theoretical concerns to set theory and model theory had a profound influence on philosophy. During the days of logical empiricism the syntactic conception of theories ruled. But from the 1960s onwards, the semantic conception of theories became dominant. Increasingly, formulating a theory meant presenting a class of models, and the language of set theory is of course the framework to do it in. This brought with it a more positive attitude to unformalised ordinary language. The latter is witnessed by the rise of philosophy of language in the 1960s.

In philosophy, some (including Tarski himself) saw Tarski's theory as a vindication of the correspondence theory of truth. Quine, in contrast, saw Tarski's theory as a vindication of the thesis that truth is only a device of quotation and disquotation. This feature of truth makes it possible for us to express infinite generalisations that could not otherwise be expressed, such as

All instances of the principle of excluded third are true.

This doctrine came to be known as *deflationism*. For Tarski, the point about the expressive power of the concept of truth was first and foremost a comment of the use of truth in meta-mathematics [18]. But Quine and his followers interpreted this as a comment on the use of truth in ordinary languages such as English. Shortly afterwards, this was taken much further. Davidson took Convention T as the cornerstone of a theory of *meaning* [8]. He took instances of Convention T to explicate the meaning of sentences of natural language (that do not contain the truth predicate).

From around 1960 onwards, philosophical logic as we now understand it emerged as a discipline. Roughly, philosophical logic can be described as the art of bringing logical theories and techniques to bear on philosophical questions. This does not mean that this was not done before, e.g. in the foundations of mathematics and in the theory of confirmation. But from the 1960s onwards this spread to most areas of philosophy, even to such core areas as metaphysics (possible worlds semantics, tense logic) and epistemology (epistemic logic, belief revision,...). And there was an expectation that, as in the foundations of mathematics, the study of paradoxes would be rewarding here, too. However, since we are dealing with philosophical concepts, it is not clear that the paradoxes will be dissolved by logic. Perhaps the best we can expect is that logical perspectives can shed light on them.[5]

Tarski's theory of truth was one of the first logical theories to be put to use in wider philosophical investigations. There is of course an indirect influence of Tarski's theory of truth (via the emergence of model theory) on the development of model-theoretic semantics for intensional logic. But there was also a direct influence of Tarski's work on truth on intensional logic. Around 1960, Tarski's theorem on the undefinability of truth was extended to an impossibility theorem in epistemic logic [21, 31]:

[5] I will return to this point in the final section.

Theorem 1 (Kaplan-Montague) *Let $\mathcal{L}_K$ be the language of arithmetic plus a primitive predicate K. Then the theory consisting of Peano Arithmetic plus the axiom*

$$K(\phi) \rightarrow \phi$$

and the rule of necessitation

$$\vdash \phi \Rightarrow \vdash K(\phi),$$

is inconsistent.

But the axiom $K(\phi) \rightarrow \phi$ and the rule of necessitation $\vdash \phi \Rightarrow \vdash K(\phi)$ seem prima facie reasonable when K is interpret as 'knowable in principle'. Later it became clear that analogues of the Kaplan-Montague paradox exist for other intensional notions, such as rational belief [38] and past-future [20].

5 Kripke

We have seen that Tarski's theory of truth is a *typed* theory of truth. It attributes the source of the liar paradox to the fact that the relevant Tarski-biconditional applies the truth predicate to a sentence (L) that itself contains the truth predicate. By restricting Convention T to sentences that do not themselves contain the truth predicate we avoid the semantic paradoxes, but the policy leaves truth attributions to sentences that themselves contain the truth predicate logically completely unconstrained.

Implicitly and indirectly, Tarski's diagnosis of the liar paradox was challenged by developments in mathematical logic since the 1930s. I am alluding here to the development of type free versions of the lambda calculus.[6] In these calculi, functions can be unproblematically applied to themselves as objects. These calculi were not as popular as typed calculi because until the late 1970s no natural model theory existed for them. For that reason (given the turn to the semantic conception of theories), the consistency of untyped calculi was not seen as a serious challenge to Tarski's theory of truth. This only changed with the rise of the theory of inductive definitions, to which we will soon turn.

In natural language we certainly do accept *some* truth attributions to sentences that themselves contain the truth predicate. For instance:

> It is true that if I were to say that Mata Hari was the first woman to climb K3, I would be saying something that is false.

Tarski was not worried by this, because he thought that natural language is inconsistent. One alternative would be to say that natural language is consistent, but we implicitly type truth predicates in natural language, so that in the displayed sentence above, the first occurrence of the truth predicate is of a higher type than the second. Many versions of this way of using Tarski's theory of truth for the study of truth in natural language have been developed. According to some such proposals, there is

[6] For details about this evolution, see [7].

indeed only one concept of truth, but this concept is an indexical notion, where the indices line up in a linear, well-founded manner as in Tarski's hierarchy [4].

Burge's theory of truth deals with the liar paradox in roughly the following manner. Consider the liar sentence L again. The truth predicate occurring in L must be 'indexed' to a particular context, which we shall call context 0. So we can say that sentence L expresses that L is not true_0. For the familiar liar argument reasons, L cannot be true_0. If L were true_0, then what it says of itself, namely that it is *not* true_0, would have to be the case, and this would yield a contradiction. But if sentence L is not true_0, then it ought to be in some sense *true* that it is not true_0. This is where, in the original argument of the liar paradox, we were led into trouble. Burge's indexical theory of truth claims that when we assert that it is true that L is not true_0, we are shifting to a new context. And it is not that we intentionally make the shift: it happens automatically. The occurrence of "true" in "it is true that S is not true_0" must be given an index different from that of "true_0"; let its index be 1. Then we have both:

- Sentence L is not true_0.
- Sentence L is true_1.

Because of the indexical shift in extension of the truth predicate between context 0 and context 1, this is not a contradiction. Thus we have a way of maintaining the uniformity of the notion of truth while at the same time helping ourselves to Tarski's hierarchy.

There is however no direct linguistic evidence that truth is an indexical notion. If we take the natural language notion of truth seriously, then it seems at first sight that there is only one such notion: it is not indexical, and it can be truthfully applied to sentences that themselves contain the truth predicate.

Suppose that we accept that natural language has exactly one truth predicate, and that the extension of the truth predicate does not shift indexically or in any other way. And suppose also that natural language (with its canons of reasoning) is consistent. Then Tarski's undefinability theorem seems to imply that natural language cannot be *semantically closed*, i.e., contain its own truth predicate. This holds as much after the semantic turn of the 1960s as before. Tarski himself stated his theorem in proof-theoretic terms, but it can just as easily be expressed in modern model-theoretic terms.

Tarski's theorem is formulated in the framework of classical logic. It is not inconceivable that if we assume a weaker logic, we can have a language that in some correspondingly weaker sense, contains its own truth predicate. Partial logic (in which some statements receive no truth value) suggests itself as an obvious candidate. Here the truth predicate is taken as a partial predicate, whereas the predicates that encode syntax (this is usually done by arithmetical predicates) are total predicates. So models of the language $\mathcal{L}_T$ can now be taken to be of the form $\mathcal{M} = \langle \mathbb{N}, \mathcal{E}, \mathcal{A} \rangle$, where $\mathcal{E}$ is the extension of the truth predicate, and $\mathcal{A}$ is the anti-extension of the truth predicate.

The reasoning of the argument of the liar paradox appears to show that both the assumption of the truth and the assumption of the falsehood of the liar sentence lead to a contradiction. Then it seems natural to suggest that the liar sentence may not have a truth value: it is neither true nor false.

This thought has been taken up at various stages and in various ways in the twentieth century. The first question is which partial logic is opted for. Some possibilities are:

- weak Kleene logic [3]
- strong Kleene logic [27]
- supervaluation [39]

Tarski's undefinability theorem entails that no partial model $\mathcal{M}$ can be such that $\mathcal{M} \models \phi \leftrightarrow T(\phi)$ for all $\phi \in \mathcal{L}_T$ [39]. But it can be hoped that there are partial models $\mathcal{M}$ such that for all $\phi \in \mathcal{L}_T$:

$$\mathcal{M} \models \phi \Leftrightarrow \mathcal{M} \models T(\phi).$$

Indeed, it was shown in 1975 that such models exist [22, 29]. Such models are called *fixed point models*.

It was at this point that Kripke entered the discussion [22]. In Kripke's theory, $\mathcal{L}_T$ is intended to be a toy model for a natural language such as English. But whereas other investigators tried to show that one fixed point model exists for one particular valuation scheme, Kripke applied the general theory of inductive definitions [32] to not only show that such fixed point models exist for a wide variety of valuation schemes for partial logic, but also to investigate their mathematical properties, such as the recursion theoretic complexity of fixed point models.

In Kripke's hands, the field of theories of truth reached a new level of mathematical sophistication. For instance, Kripke carried out calculations of recursion theoretic complexity of fixed point models: one does not find these kinds of results in philosophical logic before Kripke.[7]

There is no space to describe Kripke's model construction technique in detail here so I will just sketch the idea behind it. Kripke's fixed point models are constructed in stages. One starts with a set S_{1e} and a set S_{1a} of sentences that one takes to be definitely true, and false, respectively. Then one considers the model

$$\mathcal{M}_1 = \langle \mathbb{N}, S_{1e}, S_{1a} \rangle.$$

This model $\mathcal{M}_1$ generates a new partial model $\mathcal{M}_2$. Consider the sets:

$$S_{2e} = \{\phi \in \mathcal{L}_T \mid \mathcal{M}_1 \models \phi\};$$

$$S_{2a} = \{\phi \in \mathcal{L}_T \mid \mathcal{M}_1 \models \neg\phi\},$$

where $\models$ is the chosen valuation scheme for partial logic. Then our second partial model is

$$\mathcal{M}_2 = \langle \mathbb{N}, S_{2e}, S_{2a} \rangle.$$

The model $\mathcal{M}_2$ then generates a third model in the same way, and so on. At limit stages, one takes the unions of the extensions and the anti-extensions that have been obtained so far. If one works with a monotone valuation scheme for partial logic, then this process must eventually reach a fixed point, where no new 'definite truths' and 'definite falsehoods' are generated if one takes the next partial model. Kripke

[7] More such complexity results can be found in [5].

suggests that the stages of generation of fixed point models reflect the way in which the meaning of the truth predicate is learned by a language user who does not yet possess the concept of truth.

Cantini [7] says that in a way, "it is appropriate to say that Kripke's theory is to Tarski's as ZFC is to the theory of types." What he means by this is that Kripke has internalised Tarski's syntactic hierarchy of truth predicates into the semantics of a type free truth predicate, just as the iterative conception internalises (in its notion of rank) Russell's syntactic type theoretic hierarchy.

Kripke then classified sentences of $\mathcal{L}_T$ according to their degree of paradoxicality. The liar sentence, for instance, will be in the truth value gap in all (consistent) fixed point models. The truth teller sentence, which says of itself that it is *true*, will be gappy in the least fixed point of the valuation schemes that we have been considering, but it will have a truth value in some fixed points. Thus it has a different 'signature' of paradoxicality. Kripke also distinguished a kind of fixed point models (aside from the minimal ones) that is especially noteworthy. These are the *intrinsic fixed point models*. Intrinsic fixed point models make only sentences true that are not made false by any (consistent) fixed point model. The consistent fixed point models form a lattice under the inclusion relation, but this lattice does not contain a maximal element. However, there does exist a maximal intrinsic fixed point model. Until now, the intrinsic fixed points have not been investigated as intensively as they should perhaps be.

Kripke's article generated an exponential growth in journal articles devoted to the semantic paradoxes. As Cantini rightly says, it is difficult to overestimate the importance of Kripke's work in theories of truth [7].

A privileged role is played by the *minimal fixed point* of a valuation scheme. This model is generated when the process for building a fixed point is generated from the empty starting point, i.e., the model $\langle \mathbb{N}, \emptyset, \emptyset \rangle$. The truths of a minimal fixed point can be seen as *grounded* in the non-semantical facts. This can be seen by the fact that if $\mathcal{M}_1 = \langle \mathbb{N}, \emptyset, \emptyset \rangle$ is taken as a starting point, then the truths of the model $\mathcal{M}_2$ supervene on the arithmetical facts, since $\mathcal{M}_1$ is ignorant of any facts about truth. Then the truths of the model $\mathcal{M}_3$ supervene on those in $\mathcal{M}_2$, and so on. In this way, the supervenience on arithmetical facts is inherited by the minimal fixed point model. Some key articles devoted to the connection between Kripkean truth and groundedness are [44] and [24].

Formally, nothing changes if in Kripke's theory of truth the expression 'having no truth value' is systematically replaced by 'being both true and false'. Indeed, from the reasoning of the liar paradox Priest draws the conclusion that the liar sentence is both true and false [33]. His book has given rise to a rich literature on *paraconsistent* solutions to the semantic paradoxes. One can even go further, and in the background logic allow *both* truth value gaps and truth value gluts. If one takes a lattice-theoretic point of view to Kripkean fixed point models, then this is a natural move [40, 43].

The main problem with Kripkean theories of truth—a problem of which Kripke himself was acutely aware—is that it is questionable whether the resulting interpreted languages can really be taken to be semantically closed. To illustrate the problem, take any monotone valuation scheme for partial logic and consider any consistent fixed point model $\mathcal{M}$ of this scheme. Then it will classify the liar sentence L as

truth-valueless. But then L will a fortiori not be true. But this is expressed in $\mathcal{L}_T$ by the sentence $\neg T(L)$, and this will be just as truth-valueless as L is in $\mathcal{M}$! This is the *strengthened liar paradox*. Its investigation was already a main theme in the collection of articles [26]. Now one can say that L is untrue in a 'higher sense of being true', expressed by a new predicate T_1. But this seems to be the start of a Tarskian hierarchy of truth predicates, and avoiding such a hierarchy was one of the motivations for the type free approach to truth. Note that the indexical account is also subject to strenghened liar reasoning, for the familiar liar-reasoning shows that the sentence

This sentence is not true in any context.

cannot coherently be given a truth value. It can be argued that even paraconsistent approaches to the semantic paradox are vulnerable to strengthened liar reasoning [2].

6 Revision

As with Kripke's theory of truth, in the revision theory of truth $\mathcal{L}_T$ is intended to be a toy model for a natural language such as English. It is intended to contain all the features that are relevant for the logical properties of the notion of truth, and no more than that. The basic facts about the revision theory are described in [15], which has become the *locus classicus* on this approach.[8]

The general idea of the revision theory of truth is the following. We start with a *classical* model for $\mathcal{L}_T$. This model is transformed into a new model again and again, thus yielding a long sequence of classical models for $\mathcal{L}_T$, which are indexed by ordinal numbers. So all models that we will consider will be of the form $\langle \mathbb{N}, S \rangle$, where $\mathbb{N}$ specifies the domain of discourse and the interpretation of the arithmetical vocabulary, and S specifies the extension of the truth predicate. The official notion of truth for a formula of $\mathcal{L}_T$ is then distilled from this long sequence of models.

For simplicity, let us start with the model

$$\mathfrak{M}_0 = \langle \mathbb{N}, \emptyset \rangle,$$

the model which regards no sentence whatsoever as true. Suppose we have a model $\mathfrak{M}_\alpha$. Then the next model in the sequence is defined as follows:

$$\mathfrak{M}_{\alpha+1} = \langle \mathbb{N}, \{\phi \in \mathcal{L}_T \mid \mathfrak{M}_\alpha \models \phi\} \rangle.$$

In other words, the next model is always obtained by putting those sentences in the extension of the truth predicate that are made true by the last model that has already been obtained.

Now suppose that λ is a limit ordinal, and that all models $\mathfrak{M}_\beta$ for $\beta < \lambda$ have already been defined. Then

$$\mathfrak{M}_\lambda = \langle \mathbb{N}, \{\phi \in \mathcal{L}_T \mid \exists \beta \forall \gamma : (\gamma \geq \beta \wedge \gamma < \lambda) \Rightarrow \mathfrak{M}_\gamma \models \phi\} \rangle.$$

[8] Detailed information about the complexity of the revision theoretic notions of truth is given in [42].

In words: we put a sentence ϕ in the extension of the truth predicate of $\mathfrak{M}_\lambda$ if there is a "stage" β before λ such that from $\mathfrak{M}_\beta$ onwards, ϕ is *always* in the extension of the truth predicate. The sentences in the extension of the truth predicate of $\mathfrak{M}_\lambda$ are those that have "stabilised" to the value *True* at some stage before λ. This rule for dealing with limit stages of the revision process is called the *Herzberger rule* or the *liminf rule*.

This yields a chain of models—and a corresponding chain of extensions H_α of the truth predicate—that is as long as the chain of the ordinal numbers. Elementary cardinality considerations tell us that there must be ordinals α and β such that $\mathfrak{M}_\alpha$ and $\mathfrak{M}_\beta$ are identical: the chain of models must be periodic.

On the basis of this long sequence of models, one can then define the notion of *stable truth* for the language $\mathcal{L}_T$. A sentence $\phi \in \mathcal{L}_T$ is said to be stably true if at some ordinal stage α, ϕ enters in the extension of the truth predicate of $\mathfrak{M}_\alpha$ and stays in the extension of the truth predicate in all later models. Stable falsehood is defined in a similar way. A sentence that is neither stably true nor stably false is said to be *paradoxical*. Revision theorists have tentatively proposed to identify *truth simpliciter* with stable truth (similarly for *falsehood simpliciter*), whilst sentences that never stabilise, such as L, are classified as paradoxical.

A main objection to the revision theory of truth is that it again seems vulnerable to strengthened liar charges: I leave it to the reader to formulate this argument in detail. Another objection to the revision theory of truth is that the limit rule of the model construction does not seem to be grounded in facts about the meaning or acquisition of the concept of truth in English. Recall that in the Kripkean model generation process, nothing of interest happens at limit stages: the extensions (anti-extensions) at earlier stages are merely collected together. But in the revision theoretic process, the real action happens at the limit stages. So it is important to motivate the limit rule. Perhaps the best we can do is to say that the *liminf* rule is by far the most natural rule for resetting truth values at limit stages that we can think of.

7 Proof Theoretic Approaches

Suppose that we adopt a Wittgensteinian perspective of meaning: the meaning of an expression is given by its use, and not necessarily by an explicit definition. And suppose that we agree with Tarski that satisfying Convention T is the sole adequacy condition for a theory of truth. Then an *axiomatic* theory of truth where the instances of Convention T (restricted to sentences not containing the truth predicate) are taken as the sole non-logical axioms governing the truth predicate, seems to be all that is desired. This theory is called TB.

It turns out that TB does not prove the compositional truth axioms such as

$$\forall \phi, \psi \in \mathcal{L}_{PA} : T(\phi \wedge \psi) \leftrightarrow (T(\phi) \wedge T(\psi))$$

(where $\mathcal{L}_{PA}$ is the language of Peano-arithmetic), which do seem to be accepted in ordinary language use. But then we can of course carry out the same manoeuvre, and take the compositional truth axioms as basic principles implicitly giving the meaning of the concept of truth. The resulting theory is called $T(PA)$ (also known as CT).

The theory TB is a subtheory of $T(PA)$. Remarkably, it turns out that, in contrast to TB, the theory $T(PA)$ is not arithmetically conservative over its background theory PA (as long as we allow occurrences of the truth predicate in instances of the induction scheme). This non-conservativeness phenomenon has led to a discussion about whether a theory such as $T(PA)$ is acceptable from a deflationist perspective [19, chapter 7].

The theories TB and $T(PA)$ are typed theories of truth, and in that sense they are Tarskian in spirit. But type free theories of truth can also be constructed. Feferman has axiomatically described the inductive clauses for building Strong Kleene fixed point models [10].[9] The resulting theory KF is currently one of the most popular axiomatic theories of truth. To some extent even the supervaluation fixed points can be axiomatically described, resulting in the system VF [6]. All these theories are nonconservative over their arithmetical background theory. In fact, they are arithmetically significantly stronger than $T(PA)$.

All the foregoing type free theories are formulated within the context of classical logic. That means that strictly speaking they do not axiomatise fixed point models, but rather the models that result from fixed point models when the anti-extension of the truth predicate is taken to be the complement of the extension of the truth predicate. As a consequence, a version of the strengthened liar objection to still apply to these theories. For instance, it is not hard to see that

$$KF \vdash L \wedge \neg T(L).$$

It is possible to proof-theoretically describe the fixed point models (rather than their classical 'completions') in partial logic. If this is done for the Strong Kleene scheme, then the system PKF results [17]. But this theory is vulnerable to a criticism that has been formulated by Feferman against *most* truth theories that withdraw from full classical logic. He argues that *nothing like sustained ordinary reasoning can be carried out in partial logic* [9, p. 264].

Thus it seems that we have reached a dilemma. On the one hand, classical logic can be sustained, but then unpalatable theorems result. Indeed if we consider PA formulated in the language $\mathcal{L}_T$, we see that this theory proves

$$[L \wedge \neg T(L)] \vee [\neg L \wedge T(L)].$$

Already this theorem, which uses no non-logical truth axioms, might well give us pause. On the other hand, we may retreat from classical logic, but then we are likely to end up in a logic in which it is difficult to reason intuitively.

Instead of axiomatising a model-theoretic construction for $\mathcal{L}_T$, one can also lay down a basis B consisting of a number of minimal non-logical truth axioms that one wants to impose, and then formulate a list of optional truth axioms which one might want to add to this basis. Then all possible consistent theories extending B can be proof theoretically investigated. One such investigation was instigated in [13], and completed in [23]. Within this approach, it became clear that even certain sentences that are not self-referential can give rise to semantical paradoxes [30].

[9] The Weak Kleene variant is proof theoretically investigated in [11].

8 The Future?

I will now with some trepidation make some predictions about the future of the field of formal theories of truth. History will be my judge, so this is a perilous undertaking.

I expect questions concerning deflationism to attract *less* attention in the future than they have over the past decade or two. This debate centred on the question of non-conservativeness of truth, and it was a great catalyser for the interaction between logicians and philosophers. On the logical side interesting questions concerning conservative axiomatic theories of truth remain. But on the philosophical side the debate about deflationism somehow seems 'tired'.

We have seen that some old problems that truth theories are faced with have not been resolved. The most recalcitrant of them is probably the problem of the strengthened liar paradox. This situation should prompt us to reflect on what we expect from a theory of truth. Is it reasonable to hope that the semantic paradoxes can be *resolved* in the same sense as the set theoretic paradoxes have been resolved? Perhaps philosophical paradoxes cannot be resolved in the same way as mathematical paradoxes can be solved. Perhaps the liar paradox influences philosophical theorising about truth in a more continuous way than the Russell paradox, say, influenced set theory.

In view of this, it seems that it is time for some fundamental methodological reflection here. If the liar paradox cannot be definitively solved, then we must reflect on what we want from a theory of truth, and why. Also, we need to reflect on what kind of insight we expect from a truth theory that would meet the adequacy conditions that we will come up with.

Kripke's truth theory consists of a series of wonderfully intuitive and natural classes of models. The same can be said, to some extent at least, for the models that are produced by the revision theory of truth. It is perhaps not so easy to find a fundamentally new class of models that is simple and intuitive in a similar way—none has been found in the past few decades. But we have seen how in the past decades, models have fuelled the construction of good axiomatic systems of truth. If the supply of new models dries up, then there is a risk that intuitive axiomatic systems may be few in number in the coming decades.

In future I expect more attention to go to the logical interaction between the truth predicate and other notions that are of philosophical interest. There are some signs that this is already happening.

We have seen that many of the leading truth theories are based on a non-classical logic. Feferman's challenge for developing systems of non-classical logic in which one can reason intuitively will remain on the agenda for some years to come. Field's proposal to add a primitive conditional $\mapsto$ to $\mathcal{L}_T$ and to formulate a semantics for it goes some way to addressing this problem [12]. It is claimed that a fair amount of natural conditional reasoning can be carried out with the new conditional $\mapsto$. Field's semantics is a sophisticated combination of Kripkean ingredients and revision theoretic ingredients. In the interest of purity of theoretical motivation, an obvious question is whether a 'real conditional' can also be constructed within a purely Kripkean framework, or within a purely revision theoretic framework. Not much progress has been made on this question. Yablo has made an attempt to add a strong

conditional to the Kripkean framework [45], but his attempt was not completely successful [12, chapter 16].

The reasoning of the Kaplan-Montague paradox is structurally identical to that of the liar paradox. Yet the relation between the liar paradox and the Kaplan-Montague paradox remains ill understood. One question that is likely to receive more attention in the future is: what does a good integrated theory of type free truth and of modalities look like? In particular, it would be interesting to know whether in such a combined theory, all the blame for the paradoxes can be put on the truth predicate, in the following sense. Using the truth predicate, one can define an intensional predicate (a necessity predicate, for instance) in terms of the corresponding intensional *operator*. Can the paradoxes be avoided by restricting the logical behaviour of the truth predicate in a well-motivated way, but accepting all the standard principles of the intensional operator? The best work on this question so far may be [35].

Another question is how truth relates to notions of probability. Again, one would like to see an elegant integrated theory of type free truth and probability, where perhaps even several notions of probability can be allowed (such as physical and subjective probability), or even higher-order probability. For some initial ideas, see [25].

To conclude, I will take a slightly longer perspective and go out on a limb here. A notion of truth for a language can be seen as a special kind of *measure*. Now the notion of measure plays an important role in the theory of large cardinals—witness the notion of a 'measurable cardinal'. In other words, there may be hitherto unexplored connections between truth theory and large cardinals. This ties in with another line of research that might become more important in the years to come, to wit, the investigation of theories of typed and type free truth not for 'weak' mathematical theories such as arithmetic, but for strong theories such as set theory.[10] Our experience with truth principles added to arithmetic strongly suggests that some but not much mathematical strength can thereby be gained. It is often assumed that the same message holds when truth principles are added to set theory: some set theoretic strength can be gained, but not nearly as much as by adding large cardinal axioms. But this assumption concerning the relation between truth and set theory has not been severely tested yet.

References

1. Beall, J.C. (Ed.) (2003). *Liars and heaps*: Oxford University Press.
2. Beall, J.C. (Ed.) (2007). *Revenge of the liar. New essays on the paradox*: Oxford University Press.
3. Bochvar, D. (1981). On a three-valued logical calculus and its applications to the analysis of the paradoxes of the classical extended functional calculus. English translation from Polish by M. Bergmann. In *History and Philosophy of Logic*, *2*, 87–112.
4. Burge, T. (1979). Semantical paradox. Reprinted in Martin 1984, p. 83–117.
5. Burgess, J. (1986). The truth is never simple. *Journal of Symbolic Logic*, *51*, 663–681.

[10] See [14].

6. Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic*, *55*, 244–259.
7. Cantini, A. (2009). Paradoxes, self-reference and truth in the twentieth century. In *Handbook of the history of logic, Volume 5* (pp. 875–1013): Elsevier.
8. Davidson, D. (1984). Truth and meaning. In Davidson, D. (Ed.), *Inquiries into truth and interpretation* (pp. 17–36): Oxford University Press.
9. Feferman, S. (1984). Toward useful type-free theories I. *Journal of Symbolic Logic*, *49*, 75–111.
10. Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, *56*, 1–49.
11. Feferman, S. (1998). Axioms for determinateness and truth. *Review for Symbolic Logic*, *1*, 204–217.
12. Field, H. (2008). *Saving truth from paradox*: Oxford University Press.
13. Friedman, H., & Sheard, M. (1987). Axiomatic theories of self-referential truth. *Annals of Pure and Applied Logic*, *33*, 1–21.
14. Fujimoto, K. Classes and truths in set theory. Annals of Pure and Applied Logic, to appear.
15. Gupta, A., & Belnap, N. (1993). *The revision theory of truth*: MIT Press.
16. Halbach, V. (2011). *Axiomatic theories of truth.* Cambridge: Cambridge University Press.
17. Halbach, V., & Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic*, *71*, 677–712.
18. Hodges, W. (2008). Tarski's theory of definition. In D. Patterson (Ed.), *New essays on Tarski and philosophy* (pp. 94–132): Oxford University Press.
19. Horsten, L. (2011). *The Tarskian turn, deflationism and axiomatic truth*: MIT Press.
20. Horsten, L., & Leitgeb, H. (2001). No future. *Journal of Philosophical Logic*, *30*, 259–265.
21. Kaplan, D., & Montague, R. (1960). A paradox regained. *Notre Dame Journal of Formal Logic*, *1*, 79–90.
22. Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, *72*. Reprinted in Martin 1984, p. 53-81.
23. Leigh, G., & Rathjen, M. (2010). An ordinal analysis for theories of self-referential truth. *Archive for Mathematical Logic*, *49*, 213–247.
24. Leitgeb, H. (2005). What truth depends on. *Journal of Philosophical Logic*, *34*, 155–192.
25. Leitgeb, H. (2008). On the probabilistic convention T. *Review of Symbolic Logic*, *1*, 218–224.
26. Martin, R. (Ed.) (1970). *The paradox of the liar*: Yale University Press.
27. Martin, R. A category solution to the liar. In Martin 1970, p. 91–112.
28. Martin, R. (Ed.) (1984). *Recent essays on truth and the liar paradox*: Oxford University Press.
29. Martin, R., & Woodruff, P. (1976). On representing 'true-in-L in L. Reprinted in Martin 1984, p. 47–52.
30. McGee, V. (1985). How truth-like can a predicate be? a negative result. *Journal of Philosophical Logic*, *14*, 399–410.
31. Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflection principles and finite axiomatizability. *Acta Philosophica Fennica*, *16*, 153–167.
32. Moschovakis, Y. (1974). Elementary induction on abstract structures, North-Holland.
33. Priest, G. (1987). In contradiction, Kluwer.
34. Ramsey, F. (1926). The foundations of mathematics. *Proceedings of the London Mathematical Society*, *25*, 338–384. *Truth and probability.* In Ramsey, F. *The Foundations of Mathematics and other Logical Essays.* London: Kegan & Paul, p. 156–198.
35. Stern, J. (2012). Toward predicate approaches to modality, PhD dissertation, University of Geneva.
36. Tarski, A. (1933). The concept of truth in formalized languages. In Tarski 1983, p. 152–278.
37. Tarski, A. (1983). Logic, semantics, meta-mathematics. Translated by J.H. Woodger. Second, revised edition, Hackett.
38. Thomason, R. (1980). A note on syntactical treatments of modality. *Synthese*, *44*, 391–395.
39. van Fraassen, B. Truth and paradoxical consequences. In Martin 1970, p. 13–23.
40. Visser, A. (1984). Four-valued semantics and the liar. *Journal of Philosophical Logic*, *13*, 181–212.
41. Visser, A. (2002). Semantics and the liar paradox. In Gabbay, D., & Guenthner, F. (Eds.), *Handbook of philosophical logic*, (Vol. 10 pp. 159–245): Kluwer.
42. Welch, P.D. (2001). On Gupta-Belnap revision theories of truth, Kripkean fixed points, and the next stable set. *Bulletin of Symbolic Logic*, *7*, 345–360.
43. Woodruff, P. (1984). Paradox, truth and logic. part I: paradox and truth. *Journal of Philosophical Logic*, *13*, 213–233.
44. Yablo, S. (1982). Grounding, dependence, and paradox. *Journal of Philosophical Logic*, *11*, 117–137.
45. Yablo, S. New grounds for naive truth theory. In Beall 2003, p. 312–330.