An introduction to
Bolander, Hendricks, Pedersen (eds.): Self-Reference CSLI Publications 2006

Thomas Bolander

The book consists of an introductory chapter followed by eight original papers on self-refence. The full table of contents is given in Figure 1. Below we will give a short introduction to each of the contributing papers in the anthology. The descriptions below assume familiarity with the basic concepts and ideas of self-reference. These basic concepts and ideas are described thoroughly in the first chapter of the book.

Andrea Cantini: Fixed Point Constructions

The first paper, by Cantini, addresses the problem of finding natural formal theories of *properties* and *operations*. The first part concerns theories of properties. Since properties have no *a priori* bounds on their extensions, a property can very well apply to itself. In this way, self-reference is brought into the picture. It seems reasonable to require a formal theory of properties to satisfy the unrestricted comprehension principle, since any predicate must determine a property. However, this principle is inconsistent in a classical logical setting (Russell's paradox can be formalized in the theory). Cantini therefore considers various ways to weaken the theory in order to regain consistency. If one gives a Gentzen-style sequent calculus for the theory, the proof of its inconsistency involves a crucial application of the *contraction principle*,

$$\frac{\Gamma, \varphi, \varphi \Rightarrow A}{\Gamma, \varphi \Rightarrow A.}$$

A reasonable approach towards regaining consistency could be to try to exclude the contraction principle from the logic. Cantini shows that this gives a consistent theory, and proves a number of results concerning the obtained theory.

The second part of Cantini's paper concerns formal theories of operations. Since operations can be applied to themselves, such theories are also challenged by problems of self-reference and paradox. In fact, a version of Russell's paradox can be formalized in the naive theory of operations, demonstrating that this theory is inconsistent. Cantini shows a number of ways in which consistency can be regained by weakening the central principles. In both the property-theoretic and the operation-theoretic case, Cantini explores the border between consistency and inconsistency by considering various combinations of the relevant axioms and see whether these produce consistent systems or not.

Contents

Figure 1: Table of Contents

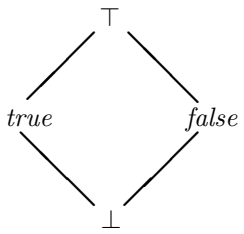

Figure 2: $\mathcal{FOUR}$

Melvin Fitting: Bilattices are Nice Things

The paper by Fitting generalizes the fixed point ideas of Kripke [1975]. Fitting's goal is to find the algebraic structure of the space in which Kripke-style truth revision operators live. He shows how the construction of Kripke fits very naturally into a setting of bilattices. The simplest non-trivial example of a bilattice is the four-valued logic called $\mathcal{FOUR}$. The three of the four values are the same as in Kleene's strong three-valued logic, and the fourth can be interpreted as a 'both-true-and-false' value (compare this to the third value, *undefined* or $\bot$, which can be interpreted as 'neither-true-nor-false'). The fourth value is denoted $\top$ (top). In the ordering put on $\mathcal{FOUR}$, $\top$ is the largest value. The ordering is illustrated in Figure 2. This ordering induces a canonical, pointwise defined, ordering on the set of $\mathcal{FOUR}$-valued interpreted languages (by a $\mathcal{FOUR}$-valued interpreted language is meant an interpreted language in which the sentences are given truth values from $\mathcal{FOUR}$). The existence of a largest truth value means that *any* set of $\mathcal{FOUR}$-valued interpreted languages will have

a least upper bound: If two languages in such a set disagree on the truth-value assigned to a certain sentence, this sentence will simply receive the truth value $\top$ (both-true-and-false) in the least upper bound (unless one of the sentences is *undefined*). The existence of least upper bounds implies that the set of $\mathcal{FOUR}$-valued interpreted languages forms a complete lattice. One can thus use all the familiar results from lattice theory on these languages, among them a standard fixed point theorem. The important thing about the set of $\mathcal{FOUR}$-valued interpreted languages is however that there are two distinct ways to order its elements so that it becomes a lattice. The first ordering is the one just considered. It is an ordering on *information*: an increase means that more sentences acquire truth values. The second ordering is induced by the ordering of $\mathcal{FOUR}$ obtained by reading Figure 2 left-right rather than up-down; in this way *true* becomes the largest truth value and *false* the smallest. The second ordering is an ordering on *truth*: an increase means that sentences become 'truer'.

In Fitting's paper, the algebraic interplay between the ordering on information and the ordering on truth is explored. He defines an interesting truth-revision operator which turns out to have nice properties with respect to both orderings, and investigates the class of fixed points of this operator (these are the aforementioned GLF-stable valuations).

Anil Gupta: Finite Circular Definitions

The third paper, by Gupta, pertains to the *theory of (circular) definitions*. This theory constitutes a central part of the *revision theory of truth* developed by Nuel Belnap and Gupta himself (Gupta and Belnap [1993]). The revision theory of truth is without doubt the most influential theory of truth and the semantic paradoxes that has been developed since the theory of Kripke [1975]. The revision theory of truth is essentially an application of the theory of definitions. The goal of the theory of definitions is to make sense of circular definitions. Since, as to be seen, truth can be considered as being a circularly defined predicate, the theory applies to theories of truth as a special case.

The theory of definitions takes its departure in the fixed point approaches. The central objects in the theory are *definitions* of the form

$$G(x_1, \ldots, x_n) =_{Df} A(x_1, \ldots, x_n, G). \tag{1}$$

Here G is a first-order predicate symbol and A is a first-order formula (possibly infinitary). The idea is that by (1) we *define* G to be the formula A. Since A can contain occurrences of G, the definition might be circular. As an example we can define a truth predicate $T(x)$ by

$$T(x) =_{Df} (x = \langle\varphi_1\rangle \wedge \varphi_1) \vee (x = \langle\varphi_2\rangle \wedge \varphi_2) \vee (x = \langle\varphi_3\rangle \wedge \varphi_3) \vee \cdots, \tag{2}$$

where $\varphi_1, \varphi_2, \varphi_3, \ldots$ is an enumeration of all the sentences in the language.[1] This is a circular definition, since the truth predicate can occur in the sentences $\varphi_1, \varphi_2, \varphi_3, \ldots$ Note that all instances of the definition can be rewritten as $T\langle\varphi_i\rangle =_{Df} \varphi_i$, corresponding to Tarski's schema T (Tarski [1935]).

[1] The predicate is defined by an infinitary formula. This does not imply that we have now moved to consider truth predicates in infinitary logic. The definition is only used to define a truth revision operator (see below), and this revision operator is still simply a mapping on standard interpreted first-order languages.

Given any definition, we can define a corresponding revision operator, called a *revision rule*, in a way similar to the truth revision operators considered in the paper by Fitting. The *revision rule* of a definition of the form (1) is the following mapping on interpreted languages

$$\Phi_G(L) = L', \text{ where } G(d_1, \ldots, d_n) \text{ is true (false) in } L' \text{ if } A(d_1, \ldots, d_n, G) \text{ is true (false) in } L.^2 \quad (3)$$

Given a definition, a *revision sequence* for it is any sequence

$$L, \Phi_G(L), \Phi_G^2(L), \Phi_G^3(L), \ldots$$

where Φ_G is the revision rule of the definition and L is an interpreted language. In the theory of definitions, revision rules are restricted to the set of totally interpreted languages. This implies that revision rules can generally not be expected to have fixed points. When a revision rule has no fixed points, none of the corresponding revision sequences will ever stabilize (there is no point after which all elements in the sequence are identical). To understand what is defined by circular definitions, one therefore instead looks for periodic patterns in the revision sequences. This leads to the concept of *finite definitions*. The definition of a predicate symbol G is called *finite* if, for all interpreted languages L, the revision sequence $L, \Phi_G(L), \Phi_G^2(L), \ldots$ is *eventually periodic*.[3] In the article, Gupta explores finite definitions in depth and proves a number of interesting results concerning their general properties.

Vann McGee: In Praise of the Free Lunch

The next paper is by McGee. Its full title is 'In Praise of the Free Lunch: Why Disquotationalists Should Embrace Compositional Semantics', and it concerns the *disquotational conception of truth*. Disquotationalists desire a theory of truth which is weak enough to plausibly be regarded as true by stipulation, yet strong enough to be useful. According to the disquotational view, the statement that a sentence is true says nothing beyond what the sentence itself conveys. Thus saying ' "snow is white" is true' is the same as simply saying 'snow is white'. This explains why the conception is called 'disquotational': Adding the words 'is true' undoes the effect of the quotation marks. Since 'P is true' and 'P' appear to be saying the same thing, it seems that having a truth predicate adds nothing at all to a language. This is however not the disquotational view. Truth is still valuable, since among other things it allows us to express generalizations about truth such as 'all theorems of the theory T are true'. Thus, truth is still expected to add expressive power to a language.

A serious problem is facing the disquotationalist view. It seems that, on this view, truth should satisfy Tarski's schema T, since given any sentence φ, the two sentences $T\langle\varphi\rangle$ and φ express the very same thing (at least if we accept the

[2] Gupta defines revision rules in a slightly different, but essentially equivalent, way.

[3] A sequence $a_1, a_2, a_3, \ldots$ is called *periodic* if there is a natural number p such that $a_i = a_{i+np}$ for all $i, n > 0$. A sequence is called *eventually periodic* if it is periodic from some point onwards. *Remark*: The definition of a finite definition given here is only equivalent to Gupta's for first-order languages.

Gödel coding $\langle\cdot\rangle$ to play the same role as quotation marks). However, as known by Tarski's theorem on the undefinability of truth (Tarski [1935]), one cannot claim schema T to hold without running into paradoxes. On the face of the paradoxes, it thus seems impossible to maintain a disquotational conception of truth.

McGee shows how the disquotational view can be saved from paradox by choosing an alternative schema to define truth. He proposes that the disquotational conception can be captured by *Tarski's inductive definition of truth*. This is the definition always used when specifying the true sentences in a first-order model. In this definition, the truth of a compound formula is defined in terms of the truth of its constituents (this makes it a *compositional semantics*). Tarski used the inductive definition to define the true sentences of first-order languages. As such, it is a definition located in a meta-language of first-order logic, but it can be reformulated as a first-order sentence in the object language. This is what McGee considers. Such a definition of truth will definitely not produce paradoxes, since a truth predicate only is introduced which applies to the original object language, that is, the language without the truth predicate. Circularity and self-reference are excluded in the same manner as in Tarski's hierarchy.

McGee argues that a positive version of the inductive definition of truth satisfies the desires of the disquotationalists. Since the inductive definition turns out to being conservative over pure logic, it can reasonably be regarded as 'true by stipulation'. At the same time, if the definition is added to first-order arithmetic, it increases the expressive power, since we will be able to prove new theorems such as the consistency of arithmetic.

Don Perlis: Theory and Application of Self-Reference

Along with Kripke and many others, Perlis suggests in his paper that truth should only be partially defined. He argues that for a sentence to be either true or false it must first have a clear enough meaning that can be measured against some criterion of truth. In some cases there is simply not a sufficiently clear separation between the meanings of 'S is true' and 'S is false' to decide that one and not the other holds. This appears for instance to be the case when S is the liar sentence. Based on these thoughts, Perlis defines a *normal order principle* by which the truth or falsity of a sentence should be determined: 'Truth (of a sentence S) is a relation between the world W, the sentence S, and the meaning m of S, where the relation has a temporal nature: W precedes S which in turn must precede m, which in turn precedes the truth (or lack thereof) of S.' It is the temporal order that is important. If the process of assigning a truth value to S following the temporal order fails, then the sentence will be neither true nor false.

Consider the normal order principle in case of sentences S of the form $T\langle\varphi\rangle$, where T is a truth predicate. For such sentences, φ is considered to be part of the world W. To determine the truth of $T\langle\varphi\rangle$ one must therefore, by the normal order principle, first look at the world to see whether φ holds, and then afterwards record this in $T\langle\varphi\rangle$. Consider the case of the liar sentence λ. To determine the truth of $T\langle\lambda\rangle$ first look at the world to see whether λ is true or not. By definition of λ, $\lambda \leftrightarrow \neg T\langle\lambda\rangle$, so what must be determined is whether

$\neg T\langle\lambda\rangle$ is true or not. This can of course not be accomplished independently of determining whether the original sentence $T\langle\lambda\rangle$ is true or not, so the temporal order of the normal order principle is violated. In other words, the process of assigning a truth value according to the principle fails, and hence the sentence $T\langle\lambda\rangle$ must be neither true nor false (it suffers from a truth-value gap).

Perlis argues that Kripke's theory of truth (Kripke [1975]) provides a formal characterization of the normal order principle. More precisely, Perlis suggests a modified T-schema as capturing much of the intuition behind the principle. However, this is not entirely unproblematic, as Perlis notes. He discusses the various problems in formalizing the normal order principle, including a discussion of what it means to *refer*. According to Perlis, reference depends inherently on there being an agent to refer, and in this sense formal sentences can never refer. Formal sentences can in particular never be self-referential, although they can still have properties leading to contradictions in close analogy to their informal counterparts.

Graham Priest: The Paradoxes of Denotation

In Priest's paper, it is discussed whether the *paradoxes of denotation* can be solved by the same methods as the other semantic paradoxes. The paradoxes of denotation are those which employ *descriptions* in an essential way. Descriptions are expressions such as 'a prime number', 'the square root of 4', and 'the least number greater than the square root of 4' (the first is an *indefinite* description and the last two are *definite*). *Berry's paradox* is one of the best known paradoxes of denotation. It arises when trying to determine the denotation of the following definite description:

> the least number that cannot be referred to by a description containing less than 100 symbols.

The contradiction is that this description containing 93 symbols denotes a number which, by definition, cannot be denoted by any description containing less than 100 symbols.

Priest considers the various standard solutions to the semantic paradoxes, and discusses whether these apply to the paradoxes of denotation. He argues that neither the Tarskian hierarchy approach nor the Kripkean truth-value gap approach give solutions to all of the denotation paradoxes. Priest also considers whether the paradoxes can be solved by appealing to either *ambiguity* or *context dependency* of descriptions, but shows that such solutions will still allow us to form strengthened versions of the denotation paradoxes.

According to Priest, all the standard paradoxes of self-reference share a common form. This includes both the set-theoretic and the semantic paradoxes. In addition, Priest subscribes to a *principle of uniform solution*: 'same kind of paradox, same kind of solution.' Given that all the paradoxes of self-reference have the same underlying structure, this principle implies that a solution to the paradoxes should be a solution to *all* of them. Since neither the Tarskian nor the Kripkean approaches provide solutions to all denotation paradoxes, these solutions do not qualify. Priest instead proposes *dialetheism* as a common solution to all paradoxes of self-reference. Dialetheism is the view that there are true contradictions (*dialetheia*), that is, there are sentences φ for which both φ

and $\neg\varphi$ are true. To capture the dialetheic view in a non-trivial formal setting, one must use a *paraconsistent logic*, that is, a logic in which a contradiction does not entail everything. Priest argues that dialetheism cannot only provide solutions to the paradoxes of denotation, but all paradoxes of self-reference, and thus it satisfies the principle of uniform solution.

Raymond M. Smullyan: Self-Reference in All Its Glory!

The paper of Smullyan consists of two parts. The first part is an entertaining and lighthearted introduction to self-reference and paradoxes. It contains several amusing self-referential anecdotes as well as some puzzles of self-reference. The second part is more mathematical, and looks at *diagonalization* and related notions in a very general setting. Diagonalization is the technical device normally used to construct self-referential sentences in formal languages. The *diagonalization* of a first-order formula $\varphi(x)$ is defined to be $\varphi\langle\varphi\rangle$, that is, $\varphi(x)$ applied to its own Gödel code. Informally, one can think of the diagonalization of a formula as being the formula applied to itself. Diagonalization is a central ingredient in achieving self-reference in first-order languages. Smullyan considers this and many other methods to achieve self-reference in a much more general setting of what he calls *sentential applicative systems*. These are abstract systems that have first-order arithmetic and a range of other mathematical systems as instances. A sentence ψ in such a system is said to be a *fixed point* of a unary predicate φ of the same system if ψ is equivalent to $\psi\varphi$ ($\psi\varphi$ means ψ applied to φ). Fixed points of sentential applicative systems can be thought of as self-referential sentences. Smullyan proves a number of general results concerning the sufficient conditions for the existence of fixed points in his systems. Since fixed points correspond to self-referential sentences, these results give some interesting insights into the general requirements that languages must satisfy in order to allow the formation of self-referential sentences.

Stephen Yablo: Circularity and Paradox

In the paper by Yablo, the following question is asked: 'Are the semantic and set-theoretic paradoxes circularity-based?' This seems to be the case when considering e.g. the liar paradox and Russell's paradox. A simple way to see this is to note that none of these paradoxes can be formulated if the universe is properly stratified to exclude circularity. In case of the liar paradox this can be accomplished by constructing a hierarchy of languages, and in case of Russell's paradox by using type theory. There are, however, semantic and set-theoretic paradoxes that cannot be escaped simply by building hierarchies. An example of such a paradox is *Yablo's paradox* (or the *ω-liar*) introduced in Yablo [1985]. This paradox can be formalized in a *descending* hierarchy of languages. Descending hierarchies differ from the usual ascending hierarchies by not having a bottom level (they are non-wellfounded). However, descending hierarchies still block circularity, so it seems that Yablo's paradox cannot be a paradox of circularity.

It is possible to construct non-circular paradoxes within set theory as well. Yablo demonstrates how a variant of Mirimanoff's paradox can be expressed in a type theory allowing negative types. This shows that a rigid type separation cannot by itself guarantee consistency. The conclusion drawn from Yablo's paradox and the variant of Mirimanoff's paradox is that to avoid the paradoxes it is not sufficient to build hierarchies—one must in addition make sure that these hierarchies are well-founded.

Yablo argues that in the setting of naive well-founded set theory, even Russell's paradox is in a certain sense not a paradox of circularity. The argument is that in well-founded set theory, where we build a cumulative hierarchy of sets, no set can be a member of itself (a set's members must always come into the hierarchy *before* the set itself). Thus for *all* sets x one must have $x \notin x$. This means that when trying to define the Russell set by

$$\{x \mid x \notin x\},$$

an attempt is actually being made to try defining the *universal set*—the set of all sets. According to Yablo, the paradox arising from assuming the existence of a universal set is in a sense not a paradox of circularity. This implies that Russell's paradox is, in the same sense, not a paradox of circularity.

In the remainder of the article, Yablo discusses the unrestricted comprehension principle and various ways to interpret it. The final conclusion is that if interpreted in the right way, unrestricted comprehension actually *does* hold and does not lead to any known paradoxes.

References

Gupta, Anil and Nuel Belnap. 1993. *The Revision Theory of Truth*. MIT Press.

Kripke, Saul. 1975. Outline of a theory of truth. *The Journal of Philosophy* 72:690–716. Reprinted in Martin [1984].

Tarski, Alfred. 1935. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* 1:261–405. English translation in Tarski [1983].

Tarski, Alfred. 1983. *Logic, semantics, metamathematics—Papers from 1932 to 1938*. Hackett Publishing Co.

Yablo, Stephen. 1985. Truth and reflection. *Journal of Philosophical Logic* 14(3):297–349.